Fall 2023

A New Method to Determine the Posterior Distribution of Coefficient Alpha

John Mart V. DelosReyes
Old Dominion University

A NEW METHOD TO DETERMINE THE POSTERIOR DISTRIBUTION OF COEFFICIENT ALPHA

by
John Mart V. DelosReyes
B.S. May 2015, Old Dominion University
M.S. December 2019, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

PSYCHOLOGY

OLD DOMINION UNIVERSITY
December 2023

Approved by:

Miguel A. Padilla (Director)

James M. Henson (Member)

Lucia M. Tabacu (Member)

ABSTRACT

A NEW METHOD TO DETERMINE THE POSTERIOR DISTRIBUTION OF COEFFICIENT ALPHA

John Mart V. DelosReyes
Old Dominion University, 2023
Director: Dr. Miguel A. Padilla

There is a focus within the behavioral/social sciences on non-physical, psychological constructs (i.e., constructs). These constructs are indirectly measured using measurement instruments that consist of questions that capture the manifestations of these constructs. The indirect nature of measuring constructs results in a need of ensuring that measurement instruments are reliable. The most popular statistic used to estimate reliability is coefficient alpha as it is easy to compute and has properties that make it desirable to use. Coefficient alpha's popularity has resulted in a wide breadth of research into its qualities. Notably, research about coefficient alpha's distribution has led to developments about its confidence intervals (CIs) and implementation through Bayesian methods with corresponding credible intervals (CrIs).

Here, a new method to implement a Bayesian coefficient alpha is proposed. This new method is built on coefficient alpha having a posterior normal distribution based on a posterior mean and variance. To assess the effectiveness of this new method, a simulation was conducted that compared it to previously established methods used to obtain coefficient alpha CI/CrIs. These include bootstrap CIs and Bayesian CrIs that are based on indirectly generating a coefficient alpha posterior though the posterior of the item covariance matrix. All CI/CrIs were assessed for 95% coverage probability. Ultimately, the new method to implement Bayesian coefficient alpha was found to be effective as it tended to have CrIs that were closer and tighter around the target of .95.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

INTRODUCTION

Measurement is a key component in the sciences as it allows for the accurate collection and description of data for phenomena of interest. For the physical sciences, measurements can generally be taken directly using the appropriate measurement instrument. For the behavioral/social sciences, measurement instruments usually cannot be directly applied to what is being measured. This is because the phenomena of interest in the behavioral/social sciences are not typically physical, but rather non-physical phenomena known as psychological constructs (hereafter referred to as constructs). Unfortunately, constructs cannot be directly observed and therefore cannot be directly measured, as such, constructs are indirectly measured with measurement instruments that record behaviors and/or responses that are manifestations of the construct of interest. Due to the indirect nature of these kinds of measurement instruments, it is essential to ensure that these instruments perform as intended by establishing their corresponding psychometric properties of reliability and validity. In terms of psychometric properties, reliability generally refers to the consistency in responding to a measurement instrument of a construct (Furr & Bacharach, 2014). There are several types of reliabilities to consider, and each has their own requirements (i.e., assumptions and settings) that warrant their use. Validity can be generally thought of as the degree in which a measurement instrument measures what it is supposed to measure. Although reliability and validity are both necessary in evaluating a measurement instrument for a construct, the focus in the current paper will be on reliability. Even so, some validity concepts will be presented as needed within the context of reliability.

Reliability is a key component in developing a valid measurement instrument. However, reliability can take on multiple forms for different purposes and these are often confused or misinterpreted in applications. This confusion about reliability can result in erroneous conclusions about the performance of a measurement instrument. There appears to be two primary sources to this confusion. First, the reliability literature is scattered over several disciplines and decades, making it difficult to fully comprehend reliability. Furthermore, over the decades, there have been mix-ups with regards to how reliability has been discussed. For example, some reliability statistics have multiple names and these are sometimes confused as being different statistics (Cho, 2016). As such, it is difficult to obtain a consistent idea to the principles of reliability via the literature. Second, there is a lack of opportunities and resources to communicate the nuances and available statistical options for establishing reliability. This can be attributed to the lack of available graduate-level training in quantitative psychology; a field specializing in measurement. In 2008, Aiken et. al. reported that 64% of 223 graduate psychology programs in the United States had courses dedicated to psychometric theory. Furthermore, as of 2020, the American Psychological Association (APA) reported that within the United States and Canada, only 88 of over 500 universities (including non-research universities) have graduate programs that specialize in quantitative psychology. As such, there may be a lack of experts in the field that can formally train graduate students or future researchers on the subject.

To help reduce the confusion about reliability, what follows is a discussion of the foundations for establishing reliability for a measurement instrument of a construct. This discussion includes a collection of currently available reliability statistics along with their appropriate use, compares the qualities of each to one another, and shows how they are related to one another. Finally, by leveraging and combining this information with Bayesian methodology, a novel Bayesian version of Cronbach’s coefficient alpha, the most popular reliability statistic, is proposed.

Before getting started, one key idea to point out is dimensionality. Dimensionality is a validity concept that refers to the number of constructs that represent a set of items (or measurement instrument). From this perspective, unidimensional indicates one construct, two-dimensional indicates two constructs, etc. Establishing dimensionality is a validity topic that is determined separately from reliability with techniques such as factor analysis. Again, reliability and validity are related, but each corresponds to different psychometric properties of a measurement instrument. Unless otherwise stated, all reliability concepts and models presented assume the items to be unidimensional.

Items, Sub-Measures, Measurement Instruments, and Composites

Although the discussion about reliability was introduced in reference to measurement instruments, it is important to clarify that Classical Test Theory, the foundation of reliability, does not make a distinction between items, sub-measures, or measurement instruments. An item refers to a singular item used to measure a construct (e.g., a question used to assess a construct). A sub-measure is a subset of items from a measurement instrument. A measurement instrument refers to a set of items used to measure a construct. For establishing reliability, it turns out that these concepts are equivalent to one another (DeVellis, 2017). As such, the term “measure” will be used when collectively referring to an item, sub-measure, and measurement instrument.

Along these lines, a composite refers to the sum of measures used to assess a construct. A composite for a set of *k* measures is defined as

$$X = \sum_{j=1}^{k} x_j \, , \tag{1}$$

where x_j is an observed score on measure *j*. In addition, the variance for the composite is defined as

$$\sigma_X^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij} = \sum_{i=1}^{k}\sigma_{ii} + 2\sum_{i=1}^{k}\sum_{j>i}^{k}\sigma_{ij} \; . \tag{2}$$

Note that items, sub-measures, or measurement instruments can all be summed with a composite. This will be made clear later when discussing Classical Test Theory, and when discussing each form of reliability to be presented. Even so, given that items are the simplest and most classical components used to establish reliability, the proceeding discussion about reliability will be in reference to items unless otherwise stated.

Classical Test Theory

The foundation for establishing reliability for an item of a construct is given within Classical Test Theory (CTT; Furr & Bacharach, 2014l Devillis, 2017). Fundamentally, CTT indicates that a respondent's observed score is based on their true score (i.e., true measurement of a construct) plus random measurement error. This idea can be expressed with the classical true score equation (CTSE)

$$x_i = \tau_i + \varepsilon_i \, , \tag{3}$$

where x_i is an observed score, τ_i is the true score, and ε_i is random measurement error for respondent *i*. In addition, it is assumed that a respondent's true score is constant and that the random measurement errors average to zero across infinite, independent, and identically repeated administrations of the same item. Infinite, independent, and identically repeated administrations indicate that the measurement error is a random variable with an expected value of zero; i.e., $E\left(\varepsilon_i\right) = 0$. As a function of the random measurement error, the observed score for each respondent is also a random variable from a propensity distribution. It can be shown that the average of the observed scores is equal to the true score for respondent *i*; i.e., $E\left(x_i\right) = \tau_i$.

Something to note is that although the most classical interpretations of CTT and CTSE were developed for items, the principles of each can be extended to sub-measures and measurement instruments. The CTSE then leads to the following three properties for a population of respondents:

- The measurement errors average to zero; i.e., $E\left(\varepsilon\right)=0$.
- The correlation between true scores and measurement errors is zero; i.e., $\rho_{\tau\varepsilon}=0$.
- The correlation between measurement errors between two separate items or two separate occasions of the same item is zero; i.e., $\rho_{\varepsilon_1\varepsilon_2}=0$.

The population properties are a direct consequence of the measurement error properties. It also turns out that the CTSE can be further conceptualized into three measurement models.

Measurement Models

There are three measurement models that dictate the appropriate reliability. These models will be presented from most restrictive to least restrictive in terms of their own assumptions. In this context, more restrictive indicates more or stronger assumptions and less restrictive indicates less or weaker assumptions. A common assumption for all the models is that the measurement errors are independent and have the same distribution. In addition, respondents will be indexed with $i=1,\ 2,\ \ldots,\ n$ and items with $j=1,\ 2,\ \ldots,\ k$.

The first measurement model to consider is the parallel model, which is the most restrictive measurement model as it assumes that items are parallel (Furr & Bacharach, 2014). Items are parallel if and only if their true scores are equal and their variability of the measurement errors are equal. This model is expressed as

$$x_{ij}=\tau_i+\varepsilon_i\,, \tag{4}$$

where x_{ij} is the observed score for respondent *i* on item *j*, τ_i is the true score for respondent *i*, and ε_i is the measurement error for respondent *i*. This model specifies that each respondent has their own true score and measurement error, but that the true score for each respondent is the same across all items (this is implied by τ not having a *j* subscript). This indicates that all the items measure the same construct with the same degree of accuracy and the same amount of error (Graham, 2006). For *k* items that follow a parallel measurement model, the following compound symmetric covariance matrix is generated (Raykov & Marcoulides, 2011):

$$\mathbf{\Sigma}_{\text{PM}} = \begin{bmatrix} \sigma_\tau^2 + \sigma_\varepsilon^2 & \sigma_\tau^2 & \cdots & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\varepsilon^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\tau^2 & \cdots & \cdots & \sigma_\tau^2 + \sigma_\varepsilon^2 \end{bmatrix}. \tag{5}$$

The second measurement model is the tau-equivalent model, which is less restrictive than the parallel model as it assumes that the items have the same true scores but with varying amounts of measurement errors (Graham, 2006; Raykov, 1997). The model is expressed as

$$x_{ij} = \tau_i + \varepsilon_{ij}, \tag{6}$$

where x_{ij} is an observed score for respondent *i* on item *j*, τ_i is the true score for respondent *i*, and ε_{ij} is the measurement error for respondent *i* on item *j*. This indicates that the items measure the same construct with the same degree of accuracy and varying amounts of measurement error.

A slightly more relaxed form of the tau-equivalent model is the essentially tau-equivalent model which allows for the true scores for the items to differ by some constant (Graham, 2006; Raykov, 1997). This model is expressed as

$$x_{ij} = \left(\alpha_j + \tau_i\right) + \varepsilon_{ij}, \tag{7}$$

where x_{ij} is an observed score for respondent *i* on item *j*, α_j is a unique constant for item *j*, τ_i is the true score for respondent *i*, and ε_{ij} is the measurement error for respondent *i* on item *j*. This indicates that the items measure the same construct with varying degrees of accuracy (i.e., potentially different true scores between items) and varying amounts of measurement error. The inclusion of the constant α_j in the essentially tau-equivalent model affects the mean of an item but not its variance nor covariance with another item. As such, for the purposes of discussing reliability, the tau-equivalent and essentially tau-equivalent model (i.e., equations 6 and 7) will be considered equivalent to one another and are hereafter collectively referred to as the (essentially) tau-equivalent model. For *k* items that follow an (essentially) tau-equivalent model, the following compound symmetric covariance matrix with heterogenous variances is generated (Raykov & Marcoulides, 2011):

$$\mathbf{\Sigma}_{\mathrm{ETM}} = \begin{bmatrix} \sigma_\tau^2 + \sigma_{\varepsilon_1}^2 & \sigma_\tau^2 & \cdots & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma_{\varepsilon_2}^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\tau^2 & \cdots & \cdots & \sigma_\tau^2 + \sigma_{\varepsilon_k}^2 \end{bmatrix}. \tag{8}$$

The final measurement model considered is the congeneric model, which is the least restrictive measurement model as it allows the true scores of the items to vary by some linear relationship (Graham, 2006; Raykov, 1997). The model is expressed as

$$x_{ij} = \left(\alpha_j + \beta_j \tau_i\right) + \varepsilon_{ij}\,, \tag{9}$$

where x_{ij} is an observed score for respondent *i* on item *j*, α_j is a unique constant for item *j*, β_j is a unique relationship for item *j*, τ_i is the true score for respondent *i*, and ε_{ij} is the measurement error for respondent *i* on item *j*. This indicates that the items measure the same construct with varying degrees of accuracy and varying amounts of measurement error. For *k* items that follow a congeneric model, the following first-order heterogeneous covariance matrix is generated for a respondent (Raykov & Marcoulides, 2011):

$$\mathbf{\Sigma}_{\mathrm{CM}} = \begin{bmatrix} \beta_1^2\sigma_\tau^2 + \sigma_{\varepsilon_1}^2 & \beta_2\beta_1\sigma_\tau^2 & \cdots & \beta_k\beta_1\sigma_\tau^2 \\ \beta_1\beta_2\sigma_\tau^2 & \beta_2^2\sigma_\tau^2 + \sigma_{\varepsilon_n}^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \beta_1\beta_k\sigma_\tau^2 & \cdots & \cdots & \beta_k^2\sigma_\tau^2 + \sigma_{\varepsilon_k}^2 \end{bmatrix}. \tag{10}$$

Determining the measurement model is crucial for estimating the appropriate reliability. Not accounting for the measurement model can result in erroneous reliability estimates (Graham, 2006; Sijtsma, 2009; Tavakol & Dennick, 2011; Walker, 2006). Graham (2006) provides a detailed method and examples for determining the measurement model for the items using structural equation modeling (SEM). A summary of this method is that the measurement models are fit for the items and are then compared to see which model fits best. Specifically, first the fit of a congeneric model is compared to the fit of an (essentially) tau-equivalent model. If the model fit is significantly different, then the items follow a congeneric model. Otherwise, the fit of the (essentially) tau-equivalent model is compared to that of a parallel model. If the models fit significantly different, then the items follow an (essentially) tau-equivalent model. Otherwise, the items follow a parallel model. The process is based on the idea that the congeneric model will always have better than or equal fit to all other models as it is the least restrictive measurement model, but the more restrictive models are justified by model parsimony.

Defining Reliability

Although the concept of reliability can generally be thought of as the consistency in responding to a measurement instrument, using CTT allows for a more technical definition of reliability by considering the relationship between true scores and observed scores (Lord & Novick, 1968; Raykov & Marcoulides, 2011). This begins with the reliability index, which is given by the correlation between true scores and observed scores and is defined as

$$\rho_{\tau x} = \frac{E\left(\tau - \mu_{\tau}\right)\left(x - \mu_{x}\right)}{\sigma_{\tau}\sigma_{x}} = \frac{\operatorname{cov}\left(\tau, x\right)}{\sigma_{\tau}\sigma_{x}} = \frac{\sigma_{\tau x}}{\sigma_{\tau}\sigma_{x}}. \tag{11}$$

Using algebra and expectation, it can be shown that

$$\rho_{\tau x} = \frac{\sigma_{\tau}}{\sigma_{x}}. \tag{12}$$

Therefore, the reliability index is the ratio of the true score standard deviation over the observed score standard deviation. Squaring the reliability index gives the reliability coefficient expressed as

$$\rho_{\tau x}^{2} = \frac{\sigma_{\tau}^{2}}{\sigma_{x}^{2}} = \rho_{xx'}. \tag{13}$$

This is the respondent's observed score variance that is explained by their true score variance of a construct and is the formal definition of reliability moving forward. Initially, this definition of reliability appears to not have much utility as applied researchers only obtain observed scores because true scores are latent and cannot be directly obtained. However, in practice, it turns out that only two observed scores are needed to obtain reliability.

For the most classical interpretations of CTT, reliability can be obtained by the correlation between the observed scores of two parallel items (i.e., measures). As defined by the parallel model (equation 4), suppose that there are two items defined as

$$\begin{aligned} x_1 &= \tau_1 + \varepsilon_1 \\ x_2 &= \tau_2 + \varepsilon_2 \end{aligned}, \tag{14}$$

where $\tau_1 = \tau_2 = \tau$ and $\sigma_{x_1}^2 = \sigma_{x_2}^2 = \sigma_x^2$. Using the definition of the correlation coefficient, the reliability for parallel items can be expressed as

$$\rho_{x_1 x_2} = \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}} = \frac{\sigma_{x_1 x_2}}{\sigma_{x_1}\sigma_{x_2}}. \tag{15}$$

Using algebra and expectation, it can be shown that the reliability for parallel items is

$$\rho_{x_1 x_2} = \frac{\operatorname{cov}(\tau, \tau)}{\sigma_x^2} = \frac{\sigma_\tau^2}{\sigma_x^2} = \rho_{xx'}. \tag{16}$$

Equation 16 is the most classical implementation of reliability provided by CTT, where $0 \leq \rho_{xx'} \leq 1$ and higher $\rho_{xx'}$ indicates better reliability (i.e., less measurement error). Although this implementation of reliability is intuitive, the scope and options of reliability have expanded over the decades. As it stands, there is no single best option when estimating reliability as there are multiple details to consider when trying to capture the consistency of a measure. These details include such things as measure implementation, underlying measure characteristics, and number of necessary measure administrations. Attempting to consider all these details at once can be daunting and lead to confusion. As such, we will further the discussion on reliability by first considering estimation methods that require two measurement instrument administrations.

Reliability Requiring Two Administrations

Test-Retest

Test-retest reliability is the earliest form of reliability and is used when a measurement instrument is administered and readministered to the same respondents after some time interval

(DeVellis, 2017; Furr & Bacharach, 2014; Raykov & Marcoulides, 2011). This form of reliability assumes that the measurement instrument follows a parallel measurement model over the administrations. Letting x_1 and x_2 be the observed scores at times 1 and 2, respectively, for the same measurement instrument (equation 14). Test-retest reliability can be estimated with the correlation of the observed scores and is expressed as

$$\rho_{\text{TR}} = \rho_{x_1 x_2} = \frac{\operatorname{cov}(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}}. \tag{17}$$

This follows the result given in equation 16 as the same measurement instrument is used in both administrations. Researchers may be interested in using test-retest reliability if they want to assess how stable a measurement instrument performs over time. In this context, this form of reliability is used to assess constructs that are stable over time or how well a person agrees with themselves over time. For example, it may be of interest to assess intelligence in adult respondents with two administrations of the same measurement instrument over a one-month interval. Intelligence is considered a stable construct and a reliable measure of intelligence should capture that stability across administrations of the measurement instrument.

Despite being simple to implement, test-retest reliability has three notable limitations (Furr & Bacharach, 2014; Raykov & Marcoulides, 2011). First, it is difficult to decide on a time interval between administrations as it depends on the construct being measured. Generally, shorter time intervals inflate reliability estimates due to measurement artifacts such as carryover or learning effects. Contrasting this, longer time intervals deflate reliability estimates. As such, if the measurement instrument measures a construct that is suspected to change over the time interval used, then test-retest reliability is not appropriate. Second, respondent maturation can be problematic. This is more common in children as they may rapidly change over the time interval due to natural changes such as cognitive development. Lastly, respondents may drop out before the second administration of the measurement instrument (i.e., attrition); creating missing data. Given that correlations are impacted by missing data, the same is true for test-retest reliability.

Alternate Forms

Building off the idea of using the same respondents twice is the idea of using two different, yet equivalent forms of a measurement instrument on the same respondents at a single point in time or short time interval. This is known as alternate forms reliability (DeVellis, 2017; Furr & Bacharach, 2014; Raykov & Marcoulides, 2011). This form of reliability assumes that the two forms of the measurement instrument follow a parallel measurement model. Letting x_{A} be the observed score of form A and x_{B} be the observed score of form B, alternate forms reliability can be estimated with the correlation of the observed scores and is expressed as

$$\rho_{\mathrm{AF}} = \frac{\mathrm{cov}\left(x_{\mathrm{A}}, x_{\mathrm{B}}\right)}{\sigma_{x_{\mathrm{A}}} \sigma_{x_{B}}} = \rho_{x_{\mathrm{A}} x_{\mathrm{B}}} . \tag{18}$$

Despite the forms of the measurement instrument being distinct from one another, the result from equation 16 still works here as the forms are equivalent to one another and can be conceptually thought of as having items that have been randomly sampled from the same content domain of the construct. Researchers may be interested in using alternate forms-reliability if they want to assess how distinct forms of a measurement instrument consistently capture the same construct.

Although alternate forms allow researchers to avoid some of the temporal-based limitations of test-retest scenarios, there are still two notable limitations (DeVellis, 2017; Furr & Bacharach, 2014; Raykov & Marcoulides, 2011). The first and most salient issue is that it is difficult to ensure that the alternative forms are equivalent (i.e., parallel) to one another. In this case, the reliability between the two forms may be negatively impacted due to how different content is presented between the two forms. Conversely, reliability between the two forms may be positively impacted if the two forms share a similar enough content. Second, respondent fatigue is also a potential source of error as the forms are responded to sequentially in the same period or short time interval. As such, reliability may be negatively impacted if the forms are too long. Additionally, the sequential administrations have potential for carryover effects as the act of taking one form of a measurement instrument can influence performance on the following form.

Test-Retest with Alternative Forms

Although test-retest and alternative forms were introduced separately, they are not mutually exclusive concepts. It is possible to combine the two concepts together to obtain reliability for alternate forms of measurement instruments that are used over some time interval. This is test-retest with alternative forms reliability (Raykov & Marcoulides, 2011). The benefit of using measurement instruments in this manner is that it allows for the evaluation of the temporal stability of a measure while minimizing things like carryover or learning effects through use of the alternative forms. This form of reliability assumes that the two forms of the measurement instrument over the time interval follow a parallel measurement model. Letting $x_{1\mathrm{A}}$ be the observed score at time 1 with form A and $x_{2\mathrm{B}}$ be the observed score at time 2 with form B, alternate forms reliability can be estimated with the correlation of the observed scores and is expressed as

$$\rho_{\mathrm{TRAF}} = \frac{\operatorname{cov}\left(x_{1\mathrm{A}}, x_{2\mathrm{B}}\right)}{\sigma_{x_{1\mathrm{A}}} \sigma_{x_{2\mathrm{B}}}} = \rho_{x_{1\mathrm{A}} x_{2\mathrm{B}}} . \tag{19}$$

Although combining test-retest with alternative forms carries some of the benefits of those methods (i.e., evaluates temporal stability and minimizing carryover effects), it also carries some of their limitations (Raykov & Marcoulides, 2011). As such, it has the same limitations as test-retest reliability: 1) difficulty in deciding on a time interval between administrations, 2) respondent maturation, and 3) respondent attrition. Like alternate forms, it may be difficult ensuring that the alternative forms are truly equivalent to one another.

Reliability Requiring One Administration

Requiring two measurement instruments and/or administrations to establish reliability can be cumbersome in applied settings. It is not always possible for researchers to obtain more than one administration of a measurement instrument for each respondent. Fortunately, the concepts of CTT can be extended such that at least lower bounds of reliability can be established from a single measurement instrument administration. Recall that CTT makes no distinction between items, sub measures, or measurement instruments (Furr & Bacharach, 2014; Raykov & Marcoulides, 2011). This allows for the relationship among measures to be used as a reliability statistic and is commonly implemented with items of a measurement instrument. This relationship is captured with the composite (i.e., sum of the measures), and is for this reason the proceeding reliabilities are sometimes referred to as composite reliabilities.

Split-Half

The first method used to obtain a reliability estimate from a single measurement instrument involves splitting a measurement instrument into two split-halves (i.e., sub-measures) after it has been administered. Obtaining a reliability estimate from these split-halves is known as split-half reliability (DeVellis, 2017; Furr & Bacharach, 2014). This is justified as it is assumed that the items are randomly sampled from the same content domain of the construct. As such, splitting the measurement instrument in half essentially creates two alternate forms. The only difference is that instead of two whole alternate forms, there are two halved alternate forms.

The use of these halved forms is important to note as it means that split-half reliability cannot be established in the same way as alternate forms reliability. This is because the reliability (i.e., correlation) between measures tends to decrease as the length of each measures decreases (i.e., the number of items decreases; Cohen, 2010; Spearman 1910). Conversely, reliability between measures tends to increase as the length of each measure increases (i.e., the number of items increases). Because splitting a measure in half creates two forms with half the items in each form, methods used to estimate reliability from split-halves account for the decrease in the number of items in each form.

Spearman Brown Prophecy Formula. The earliest method used to establish split-half reliability was the Spearman Brown (1910) prophecy formula (SBPF) defined as

$$R_{xx'} = \frac{k\rho_{xx'}}{1+\left(k-1\right)\rho_{xx'}}, \tag{20}$$

where *k* is the number of measures being examined and $\rho_{xx'}$ is the reliability (i.e., correlation) of the current pair of measures being examined. When using split-halves x_a and x_b, then $k = 2$ and the SBPF can simply be written as

$$R_{x_a x_b} = \frac{2\rho_{x_a x_b}}{1+\rho_{x_a x_b}}, \tag{21}$$

where $\rho_{x_a x_b}$ is the correlation between the split-halves.

As a reliability statistic, the SBPF works based on the properties of the composite and the assumption of a parallel measurement model. For *k* measures, the composite variance can be shown to be

$$\operatorname{var}\left(X\right)=\sum_{i=1}^{k}\operatorname{var}\left(x_i\right)+2\sum_{i=1}^{k}\sum_{i\neq j}^{k}\operatorname{cov}\left(x_i,x_j\right). \tag{22}$$

Note that this is equation 2. Furthermore, based on the definition of parallel measures, it can be shown that each measure has identical variance $\left(\sigma_x^2\right)$ and that the covariance amongst the measures is equal to the true score variance of either of them. The composite sum of the true scores is given with

$$\mathrm{T}=\sum_{i=1}^{k}\tau_i\ . \tag{23}$$

The true score variance is also common among the measures $\left(\sigma_\tau^2\right)$ due to the assumptions of the parallel measurement model. Recalling the result from equation 16, it can be shown that

$$\sigma_\tau^2=\sigma_x^2\rho_{xx'}\,. \tag{24}$$

Using these details, it can be shown that

$$\operatorname{var}\left(X\right)=\sigma_x^2\rho_{xx'}+k\left(k-1\right)\sigma_x^2\rho_{xx'}=k\sigma_x^2\left[1+\left(k-1\right)\rho_{xx'}\right] \tag{25}$$

and

$$\operatorname{var}\left(\mathrm{T}\right)=k\sigma_\tau^2+k\left(k-1\right)\sigma_\tau^2=k^2\sigma_\tau^2\,. \tag{26}$$

Combining this information allows for equation 20 (i.e., the SPBF) to serve as a reliability statistic

$$\rho_{xx'}=\frac{\operatorname{var}\left(\mathrm{T}\right)}{\operatorname{var}\left(X\right)}=\frac{k^2\sigma_\tau^2}{k\sigma_x^2\left[1+\left(k-1\right)\rho_{xx'}\right]}=\frac{k\rho_{xx'}}{1+\left(k-1\right)\rho_{xx'}}\,. \tag{27}$$

Rulon Formula. Another method used to estimate reliability for split-halves is given by Rulon (1939) and is defined as

$$R_D = 1 - \frac{\sigma_D^2}{\sigma_X^2}, \tag{28}$$

where σ_D^2 is the variance of the difference between the split-halves and σ_X^2 is the variance of the measurement instrument (i.e., the composite variance). This method was initially introduced as a shortcut to the result of the SPBF as it forgoes the correlation calculation of the split-halves. Additionally, the Rulon formula has added flexibility in that it follows an (essentially) tau-equivalent measurement model assumption. This difference in measurement model assumptions between the SBPF and Rulon Formula is important to note as it results in different outcomes between the two reliability statistics depending on how different the split-halves are from each other (Raykov & Marcoulides, 2011; Walker, 2006). Specifically, the SBPF and Rulon Formula yield different results from one another as the variances between the split-halves become more dissimilar (i.e., non-identical variances). In such cases, the Rulon Formula is preferred as it has a more relaxed assumption of (essential) tau-equivalence (Cronbach, 1951). An algebraically equivalent form to the Rulon Formula is given by Guttman (1945) as

$$R_D = G_4 = 2\left(1 - \frac{\sigma_a^2 + \sigma_b^2}{\sigma_X^2}\right), \tag{29}$$

where σ_a^2 is the variance of the split-half x_a and σ_b^2 is the variance of split-half x_b, and σ_X^2 is the composite variance.

A notable criticism about using split-halves and their related reliability statistics is that there are many ways to split a measurement instrument. In fact, for *k* items within a measurement instrument, there are

$$\frac{.5k!}{\left[(.5k)!\right]^2} \tag{30}$$

possible split-halves. This means that there is a non-unique result with regards to the reported split-half reliability of a measurement instrument (i.e., each split has its own reliability). As such, the choice in how to split the measurement instrument may impact the reported split-half reliability (Cohen, 2010). For example, simply splitting a measurement instrument in the middle is not suggested due to potential complications with regards to differences in respondent fatigue between the first and last half sections of a measurement instrument, resulting in a lower reliability estimate. To reduce such complications, an example of a better way to split a measurement instrument is to have one half be comprised of the odd items and the other half be comprised of the even items (i.e., an odd-even split). Even so, the split-half reliability of a measurement instrument can vary greatly depending on the split. This has led researchers to seek alternative methods to estimate reliability from a single measurement instrument. The most prevalent of these alternative methods are those that take advantage of the measurement instrument’s item covariance matrix.

Alternatives to Split-Halves

The item covariance matrix of a measurement instrument is useful for establishing reliability as it neatly summarizes the relationships between each item as well as the composite (DeVellis, 2017). Specifically, the diagonal components of the item covariance matrix give the variances of the items, the off-diagonal components give the relationships among the items, and the sum of all these components gives the overall variance of the measurement instrument (i.e., the composite variance). These details are important to consider as they are used to build from the framework given by the split-half approach to establishing reliability (Furr & Bacharach, 2014). To repeat, the items are considered equivalent and randomly sampled from the same content domain of the construct. However, in this case, the individual items are seen as the alternate forms. There are several ways to use the covariance matrix of a measurement instrument to establish reliability once these ideas are considered, but it should be noted that the key idea here is that the items (i.e., measures) should be related to one another as they should all be measuring the same construct. For this reason, composite reliability is also called internal consistency. With regards to this terminology, the most popular method used to establish internal consistency is coefficient alpha and is where this discussion begins.

Coefficient Alpha. Assuming that the item covariance matrix is positive definite, it can be shown that the true score variance $\left(\sigma^2_{\tau_i}\right)$ and covariance $\left(\sigma_{\tau_i\tau_i}\right)$ are related as follows

$$\sigma^2_{\tau_i} + \sigma^2_{\tau_j} \geq 2\sigma_{\tau_i\tau_j} \tag{31}$$

for each *ij* pair of true scores. This can be more generally written as

$$\sum_{i=1}^{k} \sigma^2_{\tau_i} \geq \frac{1}{k-1} \sum_{i=1}^{k} \sum_{i \neq j}^{k} \sigma_{\tau_i\tau_j} \ , \tag{32}$$

where *k* is the number of items. It can then be shown that

$$\rho_{xx'} = \alpha_c \geq \frac{\frac{k}{k-1}\left(\sum_{i=1}^{k}\sum_{j \neq i}^{k} \sigma_{x_{ij}}\right)}{\sigma^2_X} , \tag{33}$$

where $\sigma_{x_{ij}}$ is the covariance of the *ij* item pairs, and σ_X^2 is the variance of the composite. Equation 33 can be rewritten as

$$\alpha_c = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} \sigma_{x_{ii}}}{\sigma_X^2}\right), \tag{34}$$

where $\sigma_{x_{ii}}$ is the variance of item *i*. This is the familiar form of Cronbach's coefficient alpha (hereafter referred to as coefficient alpha; Cronbach, 1951).

Coefficient alpha is popular due to its ease of implementation, but its usage is only warranted when items are at most (essentially) tau-equivalent. Consider a pair of (essentially) tau-equivalent items x_{ij} and $x_{ij'}$. It can be shown through expectation that the variance of these pair of items is

$$\sigma_{x_{ij}}^2 = \sigma_{\tau_i}^2 + \sigma_{\varepsilon_{ij}}^2 . \tag{35}$$

Note that these are the elements in the covariance matrix in equation 8. Subsequently, for a respondent, it can be shown that the covariance for the items is simply σ_τ^2. Letting the sum of the true scores be denoted as

$$\mathrm{T} = \sum_{i=1}^{k} \tau_i , \tag{36}$$

the variance of the sum of the true scores can be given as

$$\begin{aligned} \sigma_{\mathrm{T}}^2 &= \mathrm{var}(\mathrm{T}) \\ \sigma_{\mathrm{T}}^2 &\geq \frac{k}{k-1}\left(\sum_{i=1}^{k}\sum_{j \neq i}^{k} \sigma_{x_{ij}}\right). \end{aligned} \tag{37}$$

See Appendix A for proof of equation 37. Altogether, this shows how equation 33 and 34 (i.e., coefficient alpha) serves as a reliability statistic when considering the measurement model assumptions. Additionally, it should be noted that coefficient alpha equals reliability only when its measurement model assumption is met. If this is not the case, coefficient alpha is only a lower bound to reliability.

Kuder-Richardson Formula 20. A reliability statistic closely related to coefficient alpha is the Kuder Richardson Formula 20 (KR20; Kuder & Richardson, 1937). The KR20 is appropriate for measurement instruments whose items only have two outcomes (i.e., are binary) and is defined as

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{\sigma_X^2}\right), \tag{38}$$

where *k* is the number of items, p_j is the proportion scoring 1 on item *j*, q_j is the proportion scoring 0 on item *j*, and σ_x^2 is the variance of the composite. The KR20 can be more generally written as

$$KR_{20} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} \sigma_j^2}{\sigma_X^2}\right), \tag{39}$$

where *k* is the number of items, σ_j^2 is the variance of item *j*, and σ_X^2 is the variance of the composite. Equation 39 shows that the KR20 is a special case of coefficient alpha. As such, the properties of coefficient alpha also extend to the KR20 with the assumption that the binary items are (essentially) tau-equivalent.

Guttman Lower Bounds. In the advent of the implementation of split-halves and composite reliability, Guttman (1945) proposed six methods to estimate a lower bound to reliability based on the (essentially) tau-equivalent measurement model. The fourth method is omitted here as it was the R_D (Guttman, 1945) for split-halves introduced earlier (equation 29).

The first method is appropriate when considering *k* measures and is defined as

$$G_1 = 1 - \frac{\sum_{j=1}^{k} \sigma_j^2}{\sigma_X^2}, \tag{40}$$

where σ_j^2 is the variance of measure *j* and σ_X^2 is composite variance. The second method directly builds from the first method and is defined as

$$G_2 = G_1 + \frac{\sqrt{\frac{k}{k-1} \sum_{i=1}^{k} \sum_{j<i}^{k} \sigma_{ij}^2}}{\sigma_X^2}, \tag{41}$$

where σ_{ij}^2 is the squared covariance of the *ij* pair of measures. The third method is a simpler alternative to the second method, and is defined as

$$G_3 = \frac{k}{k-1} G_1 . \tag{42}$$

Note that G_3 is coefficient alpha. The fifth method is defined as

$$G_5 = G_1 + \frac{\sqrt{\sigma_{ij}^{2*}}}{\sigma_X^2}, \tag{43}$$

where σ_{ij}^{2*} is the largest of the squared covariances of the *ij* pair of measures. The sixth method is based on the multiple correlation and is defined as

$$G_6 = 1 - \frac{\sum_{j=1}^{k} e_j^2}{\sigma_X^2}, \tag{44}$$

where e_j^2 is the error variance of item j from its multiple regression on the other $k-1$ items.

As lower bounds to reliability, Woodhouse and Jackson (1977) show some partial orders of the six Guttman lower bounds: $G_1 \le G_3 \le G_2$, $G_1 \le G_4$, $G_1 \le G_5$, and $G_1 \le G_6$. It is interesting to note that although G_2 outperforms G_3 (i.e., coefficient alpha), G_3 is vastly used more. This is likely due the ease in which G_3 is implemented. However, modern computational tools could change that as they have for the next reliability statistic.

Coefficient Omega. A limitation of the previously discussed reliability statistics is their measurement model assumptions. Both the parallel and (essentially) tau-equivalent measurement models are stringent and may be difficult to meet in applied settings. In contrast, a congeneric measurement model is more flexible. A reliability statistic that takes advantage of this measurement model is coefficient omega defined as

$$\omega = \frac{\left(\sum_{j=1}^{k} \lambda_j\right)^2}{\left(\sum_{j=1}^{k} \lambda_j\right)^2 + \sum_{j=1}^{k} \psi_j}, \tag{45}$$

where λ_j and ψ_j are the j^{th} factor loading and uniqueness, respectively, from a confirmatory factor analyses (CFA; McDonald, 1970). Within the context of reliability, factor loadings can be interpreted as correlations and uniqueness can be interpreted as error variance.

In general, a factor analysis is a data reduction technique based on the relationships amongst the measures and underlying common constructs (only one common construct assumed in the current context). This latter use of factor analysis is what affords coefficient omega its flexibility with the congeneric measurement model assumption. To show how the congeneric measurement model leads into reliability via coefficient omega, consider how equation 9 works within a composite (ignoring the indices for respondents for now; Raykov & Marcoulides, 2011). The decomposition of the true and observed scores of a composite here results in

$$X = \sum_{j=1}^{k} x_j = \sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau + \varepsilon_j \right) \tag{46}$$

and

$$\mathrm{T} = \sum_{i=1}^{k} \left(\alpha_j + \beta_j \tau \right). \tag{47}$$

Using the definition of reliability, it can be shown that

$$\rho_{xx'} = \frac{\mathrm{var}(\mathrm{T})}{\mathrm{var}(X)} = \frac{\mathrm{var}\left(\sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau \right) \right)}{\mathrm{var}\left(\sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau + \varepsilon_j \right) \right)} = \frac{\left(\sum_{j=1}^{k} \beta_j \right)^2 \sigma_\tau^2}{\left(\sum_{j=1}^{k} \beta_j \right)^2 \sigma_\tau^2 + \sum_{j=1}^{k} \sigma_{\varepsilon_j}^2}. \tag{48}$$

See Appendix A for proof of equation 48. Note that attempting to estimate all the parameters in equation 48 will result in a CFA model that is not identified (i.e., no unique solution for the estimated parameters). This issue can be resolved by setting one of the parameters to 1. For simplicity, it is suggested to set τ to 1 and results in

$$\rho_{xx'} = \frac{\left(\sum_{j=1}^{k} \beta_j \right)^2}{\left(\sum_{j=1}^{k} \beta_j \right)^2 + \sum_{j=1}^{k} \varepsilon_j}, \tag{49}$$

whose components have analogous interpretations to the components in equation 45. As such, coefficient omega intuitively captures the definition of reliability.

The congeneric measurement model assumption gives greater flexibility to coefficient omega with regards to determining reliability from a single measurement instrument compared to some other reliability statistics, but it still has some limitations. The most salient limitation is that coefficient omega is more difficult to compute compared to other reliability statistics (DeVellis, 2017). For example, coefficient alpha only requires the covariance matrix of the measures, whereas coefficient omega requires a CFA. This complexity also resulted in less research, development, and application of coefficient omega. However, modern computing power and software implementation has reached a point where estimating coefficient omega is viable for many applied researchers (e.g., the *psych* package in R; Revelle, 2022). In addition, modern computing power has made confidence interval research more viable for reliability statistics with respect to coefficient alpha.

Confidence Intervals for Coefficients Alpha

There are two general methods for estimating a parameter: point and interval. A point estimate uses a single value for a parameter estimate. These are useful for establishing a best guess for a parameter but do not account for sampling error (i.e., variability) about the parameter estimate. In contrast, an interval estimate consists of a range of possible values that are likely to contain the population parameter. This range is useful as it provides precision information about the parameter estimate.

A common type of interval estimate is the confidence interval (CI). A CI is formed by creating an error structure around a point estimate based on the standard error (SE) of the corresponding sampling distribution along with a confidence level (Hogg, Tanis, & Zimmerman, 2015). A confidence level indicates the consistency of a parameter estimate and is denoted as $100(1-\alpha)\%$, where α indicates the probability of a type I error. For example, a CI with $\alpha = .05$ indicates that 95% of all CIs created in the same way will contain the corresponding parameter.

CIs are useful as they provide two important pieces of information. First, they provide precision information about the parameter estimate. Specifically, narrower intervals indicate more precision and wider intervals indicate less precision. Second, they can be used for hypothesis testing (i.e., inference). These qualities are so useful that CI estimates of any statistic are broadly recommended within the literature (Beaulieu-Prevost, 2006; Benjamin et al., 2017; Wilkenson, 1999). Given the prevalence of coefficient alpha within the literature, it has seen the most development with regards to CIs.

CI Research for Coefficient Alpha

Coefficient alpha CI research has not seen much development until relatively recently. The first coefficient alpha CI was first proposed by Feldt (1965), a decade following its introduction in the 1950s. Assuming items to be parallel and normally distributed, Feldt derived a sampling distribution for coefficient alpha which allows for a coefficient alpha CI. Barchard and Haktstian (1997) investigated the performance of this CI with two Monte Carlo simulations. The first simulation considered the a) inter-item covariance (compound symmetric and spherical), b) number of items ($k = 5,\ 20$), and c) coefficient alpha ($\alpha_c = .60,\ .70, .90$). Results were based on 20,000 simulation replications and $\alpha = .05$. The second simulation considered the a) measurement model, b) number of items ($k = 5,\ 20$), c) coefficient alpha ($\alpha_c = .60,\ .70, .90$), and d) heterogeneity of item means and variances. The measurement models investigated were parallel, essentially parallel, tau-equivalent, and essentially tau-equivalent. Heterogeneity of means and variances considered combinations of item means 0, 100, or 400 with item variances 0, 10, 15, or 25. Results were based on 5,000 simulation replications and $\alpha = .05$. For both simulations, sample size was 100 and data was multivariate normal. In general, these simulations found that the CI investigated was not accurate if items were not parallel and the accompanying inter-item covariance matrix was not compound symmetric.

Following this, van Zyl et. al. (2000) showed that coefficient alpha is a maximum likelihood (ML) estimate with a normal asymptotic distribution. The derivation of this distribution required items to be normal but did require them to be parallel. Additionally, the authors showed that the results from Feldt (1965) follow the inverse hyperbolic transformation used to make a coefficient alpha normal when items are parallel. The authors compared their asymptotic normal distribution for coefficient alpha to the exact *F*-distribution proposed by Feldt via simulation when items were parallel. The authors checked for parity between the distribution under combinations of a) sample size ($n = 10,\ 30$, 50, 100), b) number of items ($k = 3,\ 6$, 14), and c) coefficient alpha ($\alpha_c = .7224,\ .8514$, .6191). Results were based on 500 simulation replications. It was found that that the asymptotic normal distribution fit the exact *F*-distribution well based on having similar probability densities, means, and variances. This was beneficial as it potentially let coefficient alpha CIs be estimated without assuming items being parallel. Duhachek and Iaobucci (2004) would build off these findings to propose a normal theory (NT) CI for coefficient alpha. This NT CI was compared to alternative reliability CIs proposed by Feldt (1965), Hakstain and Whalen (1976), Nunnally and Bernstein (1994), Lord and Novick (1968), and Charter (2000) via a Monte Carlo simulation. The simulation conditions were a) sample size ($n = 30,\ 50$, 100, 200), b) number of items ($k = 5,\ 7$), and c) mean item correlation ranging from 0.4 to 0.7. Items were normal and considered cases that were unidimensional or two dimensional. Results were based on 1,000 simulation replications and $\alpha = .05$. It was found that the NT CI outperformed all other CIs investigated across all simulation conditions.

Although CI development for coefficient alpha made estimation less reliant on assumptions of item parallelism, there was still restrictions with regards to the assumption of items being continuously normal. This is an issue as measurement instruments tend to use Likert-type items that are not continuous and may not be normal. To overcome this limitation, Yuan, Guarnaccia, and Hayslip (2003) proposed a bootstrap CI for coefficient alpha. The bootstrap is a statistical technique that uses resampling of data (i.e., bootstrap samples) to generate an empirical sampling distribution (ESD) of a statistic from which inferences can be made. It is a potentially powerful tool that requires no distributional assumptions but is computationally intensive. Common bootstrap CIs based on the percentiles of the ESD include the percentile bootstrap (PB) CI and the bias-corrected and accelerated (BCA) CI. Additionally, the authors proposed a new asymptotic distribution free (ADF) CI for coefficient alpha. The authors compared the NT, ADF, PB, and BCA CIs on data collected from the Hopkins Symptom Checklist (HSCL; Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974). The HSCL consists of five dimensions with 58 4-point Likert-type items. The sample size of this study was 419 people, all bootstrap CIs were based on 1,000 bootstrap samples, and all CIs were based on $\alpha = .10$. The results indicated that the PB and BCa were the most accurate, followed by the ADF, and NT CIs; respectively. However, the difference in performance between the bootstrap and ADF CIs were all within three decimals places (Maydeu-Olivares, Coffman, & Hartmann, 2007).

As the ADF CI performed similarly to the bootstrap CIs while also being less computationally intensive, there was interest in further investigating its properties. Maydeu-Olivares et. al. (2007) investigated the ADF CI and how it compared to the NT CI in two simulations. Interest here was on how the ADF and NT CIs performed with non-normal items. The first simulation investigated parallel items with a) sample size ($n = 50,\ 100, 200, 400$), b) number of items ($k = 5,\ 20$), c) common correlation ($\rho = .16,\ .36, .64$), and d) item type. The second simulation investigated congeneric items with a) sample size ($n = 50,\ 100, 400, 1000$), b) number of items ($k = 7,\ 14, 21$) and c) item type. For both simulations, item type consisted of six combinations of Likert-type items with the following varying skewness and kurtosis:

- 2 categories ($\text{skewness} = 2.67,\ \text{kurtosis} = 5.11$)
- 2 categories ($\text{skewness} = 1.96,\ \text{kurtosis} = 1.84$)
- 2 categories ($\text{skewness} = 0.41,\ \text{kurtosis} = -1.83$)
- 5 categories ($\text{skewness} = 0,\ \text{kurtosis} = -1.83$)
- 5 categories ($\text{skewness} = 0,\ \text{kurtosis} = -0.5$)
- 5 categories ($\text{skewness} = 0.98,\ \text{kurtosis} = -0.2$)

Results were based on 1,000 replications and $\alpha = .05$. The following results were found similar for both parallel and congeneric items. When items were distributed close to normal, the NT CI had adequate coverage probability with sample sizes $n \ge 50$. However, as items became more non-normal, the ADF CI outperformed the NT CI. Generally, when $n \ge 100$ and items are non-normal, the ADF CI performed well.

As more coefficient alpha CIs have developed, so did the interest in finding the best performing CIs. Romano, Kromrey, and Hibbard (2010) compared the performance of eight coefficient alpha CIs with a Monte Carlo simulation. The eight CI methods were a) the Bonett z-transformation (2002), b) Fisher z-transformation (1950), c) Hakstian and Whalen z-transformation (1976), d) Feldt *F*-distribution (1965), e) Duhachek and Iacobucci NT (2004),

f) Koning and Franses *F*-distribution (2003), g) Koning and Franses asymptotic, and h) Maydeu-Olivares et. al. ADF (2007). The study investigated a) coefficient alpha ($\alpha_c = .50,\ .70,\ .90$), b) sample size ($n = 10,\ 50,\ 100,\ 1000$), and c) number of items ($k = 5,\ 10,\ 20,\ 40$). All items were dichotomous and generated from a 3-parameter item response theory model. Results were based on 10,000 simulation replications and $\alpha = .01,\ .05,\ .10$. The results indicated that the Fisher z-transformation and Bonett z-transformation CIs had the best coverage probability performance. Interestingly, in contrast to previous research, it was found that the ADF CI had the poorest coverage probability performance.

From here, there was further interest in using the bootstrap to estimate coefficient alpha CIs. Padilla, Divers, and Newton (2012) proposed investigating bootstrap coefficient alpha CIs via simulation and seeing how they compared to previously investigated CIs with regards to their performance with non-normal data. The assessed bootstrap CIs were the normal theory bootstrap (NTB), PB, and BCA CIs. The assessed non-bootstrap CIs were the Fisher z-transformation (1950), Bonett z-transformation (2002), Duhachek and Iacobucci NT (2004), and Maydeu-Olivares et. al. ADF (2007) CIs. The simulation investigated the effects of a) number of Likert-type items ($k = 5,\ 10,\ 15,\ 20$), b) correlation type, c) number of item responses ($c = 2,\ 3,\ 5,\ 7$), d) distribution type, and e) sample size ($n = 50,\ 100,\ 150,\ 200,\ 250,\ 300$). For correlation type, three types of item correlation matrices were used. The first two correlation matrices corresponded to a parallel item model with a common correlation of $\rho = .30$ or .56, respectively. The third correlation matrix corresponded to a congeneric item model. For distribution type, three different distribution types were investigated based on number of item responses. When there were only 2 responses, the distributions were:

- Type 1: $\text{skewness} = 0$ and $\text{kurtosis} = -2$
- Type 2: $\text{skewness} = 1.70$ and $\text{kurtosis} = 0.88$
- Type 3: $\text{skewness} = 0.41$ and $\text{kurtosis} = -1.83$

When number of item responses 3, 5, or 7, the distributions were:

- Type 1: $\text{skewness} = 0$ and $\text{kurtosis} = 0$
- Type 2: $\text{skewness} = 0$ and $\text{kurtosis} = 0.88$
- Type 3: $\text{skewness} = 0.97$ and $\text{kurtosis} = -0.20$

The bootstrap CIs used 2,000 bootstrap samples. Results for the bootstrap CIs were based on 1,000 replications and results for the non-bootstrap CIs were based on 25,000 replications with $\alpha = .05$ in both cases. In contrast to previous research, it was found that the Fisher z-transformation, Bonett z-transformation, and NT CIs had the most instances of inadequate performance. However, the Bonett z-transformation and NT CIs had generally adequate performance if items were normally distributed or had little skewness. Additionally, it was found that the NTB CI was the best performing as it had adequate performance across most of the simulation conditions; it had unacceptable coverage in only 4 of the 864 investigated conditions.

Bayesian Coefficient Alpha

Up to this point, the discussion about coefficient alpha has implicitly been within the context of traditional statistics (i.e., frequentist statistics). However, developments in recent years have enabled exploration into how to approach coefficient alpha from a Bayesian perspective.

Bayesian Methodology

A primary goal in using statistics is to obtain an understanding for a phenomenon through a subset of information. This can be seen with how a characteristic of a population θ is approximated by a sample (subset) of that population y. In traditional statistics, θ is a constant and the only information available to approximate θ is through the sample y. However, in Bayesian statistics, all unknowns are considered random variables by way of Bayes theorem

$$\pi\left(\theta|y\right)=\frac{p\left(y|\theta\right)p\left(\theta\right)}{\int_{\Theta}p\left(y|\theta\right)p\left(\theta\right)d\theta}\propto p\left(\theta|y\right)p\left(\theta\right), \tag{50}$$

where $\pi\left(\theta|y\right)$ is the posterior distribution, $p\left(\theta\right)$ is the prior distribution, $p\left(y|\theta\right)$ is the data likelihood function and $\int_{\Theta}p\left(y|\theta\right)p\left(\theta\right)d\theta$ is a normalizing constant. Here, both θ and y are uncertain, but after obtaining the data y (i.e., $p(y|\theta)$), this information can be used to decrease the uncertainty of θ. Furthermore, previous information or beliefs about θ (i.e., $p(\theta)$) can also influence the uncertainty about θ. Finding the change in this uncertainty of θ (i.e., $\pi(\theta|y)$) is the purpose of Bayesian inference.

Bayesian statistics can be more challenging to implement than traditional statistics but yield great benefits. In Bayesian statistics, information about θ is impacted by data and prior information about θ. This can be useful in cases where this sample size is limited but prior research about the phenomenon is available. Additionally, Bayesian credible intervals (CrIs) have a more intuitive interpretation than their traditional counterparts (i.e., frequentist CIs). For example, a 95% CI is interpreted as 95% of all similarly constructed CIs contain the parameter of interest θ. In contrast, a 95% CrI is interpreted as the parameter θ having a 95% chance of being within the bounds of the of CrI given the data. Although these qualities benefit many statistics, the focus here will be on the most popular reliability statistic, coefficient alpha.

Previous Developments for Bayesian Coefficient Alpha

Padilla and Zhang (2011) proposed an initial Bayesian coefficient alpha. The authors noticed that coefficient alpha is completely determined by the item covariance matrix. As such, the posterior coefficient alpha can be obtained by using the posterior of the item covariance matrix. Starting with a multivariate normal distribution for the data, the corresponding posterior distribution is given by

$$\pi\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\middle|\boldsymbol{y}\right)\propto p\left(\boldsymbol{y}\middle|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=p\left(\boldsymbol{y}\middle|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\mu}\middle|\boldsymbol{\Sigma}\right)p\left(\boldsymbol{\Sigma}\right). \tag{51}$$

Separate priors are necessary for the covariance matrix and mean vector. Using the following conjugate priors for the covariance matrix

$$\boldsymbol{\Sigma}\sim W^{-1}\left(d_0,\boldsymbol{SS}_0\right) \tag{52}$$

and mean vector

$$\boldsymbol{\mu}\middle|\boldsymbol{\Sigma}\sim N\left(\boldsymbol{\mu}_0,\frac{1}{n_0}\boldsymbol{\Sigma}\right), \tag{53}$$

we can obtain the corresponding posterior distributions for the covariance matrix

$$\boldsymbol{\Sigma}\middle|\boldsymbol{y}\sim W^{-1}\left(n+d_0,\left(n-1\right)\boldsymbol{S}+\boldsymbol{SS}_0+\frac{nn_0}{n+n_0}\left(\overline{\boldsymbol{y}}-\boldsymbol{\mu}_0\right)\left(\overline{\boldsymbol{y}}-\boldsymbol{\mu}_0\right)'\right) \tag{54}$$

and mean vector

$$\boldsymbol{\mu}\middle|\left(\boldsymbol{\Sigma},\boldsymbol{y}\right)\sim N\left(\frac{1}{n+n_0}\left(n\overline{\boldsymbol{y}}+n_0\boldsymbol{\mu}_0\right),\frac{1}{n+n_0}\boldsymbol{\Sigma}\right), \tag{55}$$

where W^{-1} denotes an inverted Wishart distribution, d_0, $\boldsymbol{SS}_0$, $\boldsymbol{\mu}_0$, and n_0 are hyperparameters selected by the analyst, and $\overline{\boldsymbol{y}}$ and $\boldsymbol{S}$ are the mean vector and covariance matrix estimated from the data, respectively. By simulating $t = 1, 2, \ldots,$ T values from equation 54 as $\boldsymbol{\Sigma}^{(t)} \left| \boldsymbol{y} \right.$, the estimation of the coefficient alpha posterior can be obtained as

$$\alpha_c^{(t)} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_{ii}^{(t)}}{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma_{ij}^{(t)}}\right), \tag{56}$$

where $\sigma_{ii}^{(t)}$ and $\sigma_{ij}^{(t)}$ are elements from $\boldsymbol{\Sigma}^{(t)} \left| \boldsymbol{y} \right.$. Note that $\boldsymbol{\mu} \left| \left(\boldsymbol{\Sigma}, \boldsymbol{y} \right) \right.$ is not necessary as $\alpha_c^{(t)}$ only relies on $\boldsymbol{\Sigma} \left| \boldsymbol{y} \right.$. A Bayesian coefficient alpha is then obtained from the posterior mean as

$$\alpha_b = E\left(\alpha_c \left| \boldsymbol{\Sigma}, \boldsymbol{y} \right.\right) \tag{57}$$

or as the posterior median as

$$p\left(\alpha_c \left| \boldsymbol{y} \le \alpha_{b,m} \right.\right) = p\left(\alpha_c \left| \boldsymbol{y} \ge \alpha_{b,m} \right.\right) \ge 1/2. \tag{58}$$

The authors assessed their proposed Bayesian coefficient alpha by the performance of two types of CrIs via simulation (Padilla & Zhang, 2011). One of the investigated CrIs was based on the lower $\alpha / 2$ and $1 - \alpha / 2$ percentiles of the posterior, where α was the type I error rate. The second investigated CrI was normal theory-based and given by

$$\hat{\alpha}_b \pm z_{\alpha/2} SE\left(\hat{\alpha}_b\right), \tag{59}$$

where $z_{\alpha/2}$ is a value taken from the standard normal distribution based on α and $SE\left(\hat{\alpha}_b\right)$ is the α_b standard error (i.e., standard deviation). The simulation investigated the a) number of items ($k = 5, 10, 15, 20$), b) mean item correlation ($\bar{\rho} = .1667, .2208, .3103$), and c) sample size

($n = 50,$ 100, 150, 200, 250, 300). The mean item correlation resulted in coefficient alphas ranging from .50 to .90 and is summarized in Table 1.

Table 1

Population Coefficient Alpha for Items by Mean Item Correlations (Padilla & Zhang, 2011)

Items	Mean Item Correlation		
	.1667	.2208	.3103
5	.5001	.5862	.6923
10	.6667	.7392	.8182
15	.7500	.8095	.8709
20	.8000	.8500	.9000

For each simulation condition, 1,000 replications were obtained. Priors used for the Bayesian coefficient alpha were set to be non-informative. It was found that the percentile based CrIs had better performance than the normal theory-based CrIs. Generally, the normal theory-based CrI struggled when $\bar{\rho} = .3101$ and had further issues as number of items increased. However, this limitation is minimized if $n \geq 250$. The percentile based CrIs had unacceptable performance if sample size was $n = 50$ and if number of items was $k \geq 15$ but had generally acceptable performance if $n \geq 100$.

Novel Bayesian Coefficient Alpha

Although the previous study provides a general framework for implementing a Bayesian coefficient alpha, it is based on the posterior covariance matrix. As such, the method proposed by Padilla and Zhang (2011) requires a prior for the covariance matrix. However, the covariance matrix is not often included when reporting coefficient alpha and it is difficult to derive the covariance matrix from just coefficient alpha. Unfortunately, for this version of Bayesian coefficient alpha, applications requiring an informative prior can be difficult to implement. To overcome this, a new posterior of coefficient alpha is proposed.

First, under the assumption of normally distributed items, van Zyl et. al. (2000) showed that

$$\hat{\alpha}_c \sim N\left(\alpha_c, \frac{\sigma_\alpha^2}{n}\right), \tag{60}$$

where

$$\sigma_\alpha^2 = \left(\frac{k}{k-1}\right)^2 \left[\frac{2}{\left(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}\right)^3}\right] \left[\left(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}\right)\left(tr\left(\boldsymbol{\Sigma}^2\right) + tr\left(\boldsymbol{\Sigma}\right)^2\right) - 2tr\left(\boldsymbol{\Sigma}\right)\left(\mathbf{1}'\boldsymbol{\Sigma}^2\mathbf{1}\right)\right], \tag{61}$$

σ_α^2 is the coefficient alpha variance, *n* is the sample size, *k* is the number of items, and $\boldsymbol{\Sigma}$ is the item covariance matrix. The variance estimate $\left(S_\alpha^2\right)$ can be obtained by replacing $\boldsymbol{\Sigma}$ with the sample estimate $\boldsymbol{S}^2$.

Second, let $\boldsymbol{y} = y_1, \ \ldots, \ y_n$ and using Bayes' rule for the normal distribution, the posterior for coefficient alpha is

$$\pi\left(\alpha_c, \sigma_\alpha^2 \middle| \boldsymbol{y}\right) = \frac{p\left(\boldsymbol{y} \middle| \alpha_c, \sigma_\alpha^2\right) p\left(\alpha_c, \sigma_\alpha^2\right)}{p\left(\boldsymbol{y}\right)}. \tag{62}$$

Note that the joint distribution for quantities can be expressed as a product of a conditional probability and a marginal probability as

$$p\left(\alpha_c, \sigma_\alpha^2\right) = p\left(\alpha_c \middle| \sigma_\alpha^2\right) p\left(\sigma_\alpha^2\right). \tag{63}$$

Then dropping the denominator in equation 62 gives

$$\pi\left(\alpha_c,\sigma_\alpha^2\middle|\boldsymbol{y}\right)\propto p\left(\boldsymbol{y}\middle|\alpha_c,\sigma_\alpha^2\right)p\left(\alpha_c\middle|\sigma_\alpha^2\right)p\left(\sigma_\alpha^2\right). \tag{64}$$

To have simpler posterior calculations, conjugate priors will be used for the parameter of interest. If σ_α^2 is known, the conjugate prior for α_c is

$$\alpha_c\mid\sigma_\alpha^2\sim N\left(\alpha_0,\frac{\sigma_\alpha^2}{\kappa_0}\right). \tag{65}$$

Here, α_0 and κ_0 can be thought of as the coefficient alpha and sample size, respectively, from a set of prior observations. This leaves a prior for σ_α^2 to be determined. The σ_α^2 requires a distribution that has support for $(0,\infty)$. The inverse-gamma (IG) distribution family meets this requirement for the normal variance. As such, the conjugate prior for σ_α^2 is

$$\sigma_\alpha^2\sim IG\left(\frac{\upsilon_0}{2},\frac{\upsilon_0\sigma_0^2}{2}\right). \tag{66}$$

Here, υ_0 and σ_0^2 can be thought of as the sample size and variance, respectively, of a set of prior observations.

Suppose that a set of items are normally distributed, then the sampling distribution is

$$\boldsymbol{y}\middle|\alpha_c,\sigma_\alpha^2\sim NID\left(\alpha_c,\sigma_\alpha^2\right). \tag{67}$$

Using the conjugate priors in equations 65 and 66, respectively, the posterior for coefficient alpha is then

$$\alpha_c \left| \boldsymbol{y}, \sigma_\alpha^2 \sim N\left(\alpha_n, \frac{\sigma_\alpha^2}{\kappa_n}\right)\right., \tag{68}$$

where

$$\alpha_n = \frac{\left(\kappa_0 / \sigma_\alpha^2\right)\alpha_0 + \left(n / \sigma_\alpha^2\right)\hat{\alpha}_c}{\kappa_0 / \sigma_\alpha^2 + n / \sigma_\alpha^2} = \frac{n\hat{\alpha}_c + \kappa_0\alpha_0}{\kappa_n}, \tag{69}$$

$\kappa_n = n + \kappa_0$, and $\hat{\alpha}_c$ is the sample coefficient alpha. The posterior distribution of σ_α^2 can be obtained by integration over α_c as

$$\begin{aligned} p\left(\sigma_\alpha^2 | \boldsymbol{y}\right) &\propto p\left(\sigma_\alpha^2\right) p\left(\boldsymbol{y} \middle| \sigma_\alpha^2\right) \\ &= p\left(\sigma_\alpha^2\right) \int p\left(\boldsymbol{y} \middle| \alpha_c, \sigma_\alpha^2\right) p\left(\alpha_c \middle| \sigma_\alpha^2\right) d\alpha \end{aligned}. \tag{70}$$

Completing the integral in equation 70 results in

$$\sigma_\alpha^2 \left| \boldsymbol{y} \sim IG\left(\frac{\upsilon_n}{2}, \frac{\upsilon_n \sigma_n^2}{2}\right)\right., \tag{71}$$

where

$$\sigma_n^2 = \frac{1}{\upsilon_n}\left[\left(n-1\right)S_\alpha^2 + \upsilon_0\sigma_0^2 + \frac{\kappa_0 n}{\kappa_n}\left(\hat{\alpha}_c - \alpha_0\right)^2\right], \tag{72}$$

$\nu_n = n + \upsilon_0$, and S_α^2 is the sample variance from equation 61. With non-informative priors, the coefficient alpha posterior is completely determined by the data as

$$\alpha_c \left| \boldsymbol{y}, \sigma_\alpha^2 \sim N\left(\hat{\alpha}, \frac{\sigma_\alpha^2}{n}\right)\right. \tag{73}$$

and the variance as

$$\sigma_{\alpha}^{2}|\boldsymbol{y} \sim IG\left(\left(n-1\right),\left(n-1\right)S_{\alpha}^{2}\right). \tag{74}$$

Assuming a measurement instrument from prior observations (n_{pr}) with *k* parallel items, using prior information will require the α_0 and σ_0^2 priors. Obtaining the α_0 prior is easy as coefficient alpha is readily reported in the literature for a measurement instrument. However, obtaining σ_0^2 is a little more challenging as it requires the common item variance (σ_x^2). Note that the common variance $\left(\sigma_x^2\right)$ is not the composite variance $\left(\sigma_X^2\right)$. Recall that a parallel model has a compound symmetric covariance matrix for the items (Equation 5), which can be concisely written as

$$\hat{\boldsymbol{\Sigma}}_{PM} = \hat{\sigma}_x^2\left[\hat{\rho}_{xx'}\mathbf{11}'+\left(1-\hat{\rho}_{xx'}\right)\mathbf{I}\right], \tag{75}$$

where $\hat{\sigma}_x^{\mathbf{2}}$ is the estimated common variance of the items and $\hat{\rho}_{xx'}$ the estimated common reliability among the items (van Zyl et al., 2000). Again, the parallel model implies that items have the same variance and reliability. Using these details, coefficient alpha can be written as

$$\hat{\alpha}_c = \frac{k\hat{\sigma}_x^2\hat{\rho}_{xx'}}{1+\left(k-1\right)\hat{\rho}_{xx'}}. \tag{76}$$

Then, equation 76 can be solved for the common reliability as

$$\hat{\rho}_{xx'} = \frac{\hat{\alpha}_c\hat{\sigma}_x^2}{k+(1-k)\hat{\alpha}_c\hat{\sigma}_x^2}. \tag{77}$$

Taken all together, under the parallel measurement model, given the coefficient alpha estimate ($\hat{\alpha}_c$) and the common item variance estimate ($\hat{\sigma}_x^2$) allows for one to obtain the common reliability estimate ($\hat{\rho}_{xx'}$) and reconstruct the covariance matrix $\left(\boldsymbol{\Sigma}\right)$ using equation 75. Again, note that the common variance is not the composite variance. From here, $\hat{\sigma}_{\alpha}^{2}$ can be computed from equation 61. Although the common item variance is not commonly reported in the literature, variances for the items are. As such, a simple way to compute the common variance is to average the item variances as

$$\hat{\sigma}_{x}^{2} = \frac{1}{k}\sum_{j=1}^{k}\hat{\sigma}_{j}^{2} . \tag{78}$$

Then the priors can be set as $\alpha_0 = \hat{\alpha}_c$, $\sigma_0^2 = \hat{\sigma}_{\alpha}^{2}$, and $\nu_0 = \kappa_0 = n_{pr}$.

For (essentially) tau-equivalent and/or congeneric models, the method previously discussed cannot be used for obtaining the σ_0^2 prior. The reason is that for either of these models, the item covariance matrix cannot be written as in equation 75. As such, for either of these models, a σ_0^2 prior will need to be provided or the item covariance from which the prior can be obtained using equation 61.

Interval Estimation

The posterior distribution $\pi\left(\alpha_c \middle| \boldsymbol{y}, \sigma_{\alpha}^{2}\right)$ can then be summarized for inferences. Here, three different Bayesian CrIs are of interest. First, a NT CrI can be obtained with

$$\hat{\alpha}_c \pm z_{\alpha/2} SE\left(\hat{\alpha}_c\right), \tag{79}$$

where $SE\left(\hat{\alpha}_c\right)$ is the standard deviation of $\pi\left(\alpha_c \middle| \boldsymbol{y}, \sigma_{\alpha}^{2}\right)$. Second, percentile-based CrI can be obtained by using the lower $\alpha/2$ and upper $1-\alpha/2$ percentiles from $\pi\left(\alpha_c \middle| \boldsymbol{y}, \sigma_{\alpha}^{2}\right)$. Third, the highest probability density (HPD) CrI can be obtained by letting a subset of the parameter space for $\hat{\alpha}_c$, $c \subset A$, be defined as

$$c = \left\{ \hat{\alpha}_c : \pi\left(\alpha_c \mid \boldsymbol{y}, \sigma_\alpha^2\right) \geq h \right\}, \tag{80}$$

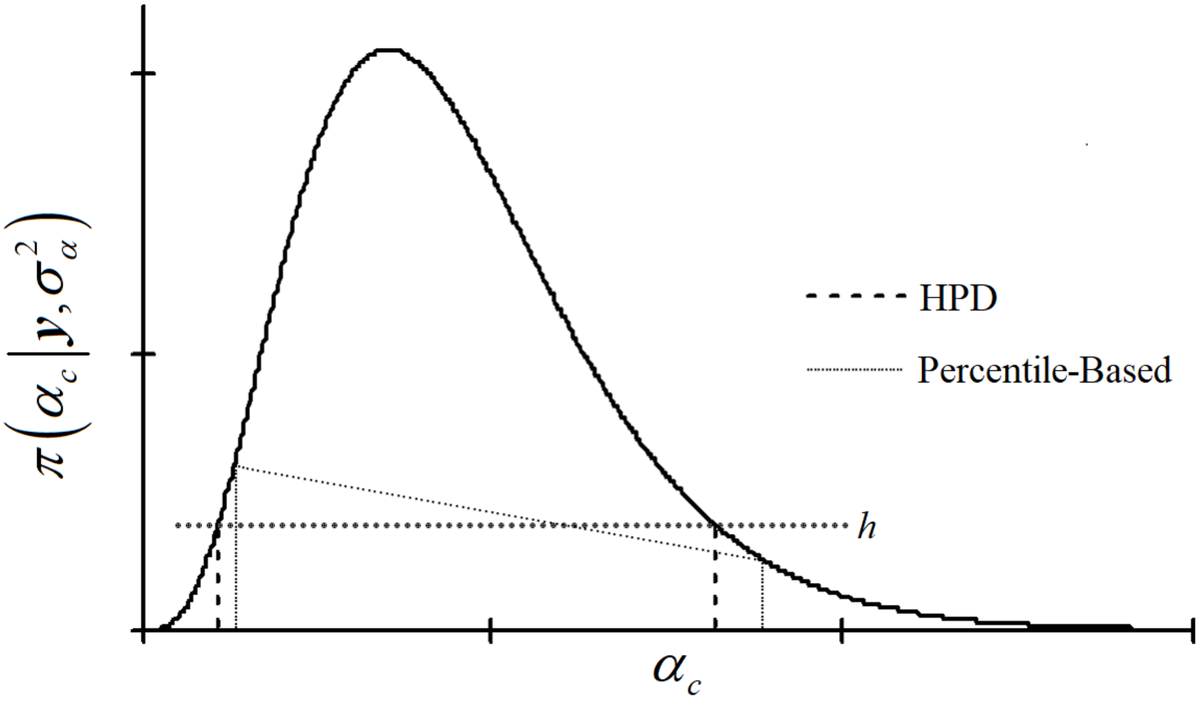
where *h* is the largest number such that

$$\int_h \pi\left(\alpha_c \mid \boldsymbol{y}, \sigma_\alpha^2\right) d\alpha_c = 1 - \alpha . \tag{81}$$

The HPD CrI has unique attributes that make it potentially useful to researchers. For a unimodal distribution, HPD CrI is defined as the narrowest interval that contains the specified probability of the confidence level $\left(1-\alpha\right)$. Here, *h* can be thought of as a horizontal line that shifts down through the distribution until its intersections with the distribution capture the $1-\alpha$ region. This results in a region that is projected upon the *x*-axis as an interval. Consequently, all *x*-axis points inside the interval have higher probability than any *x*-axis point out the interval. Additionally, the bounds of this interval will have equal posterior probability to one another. Other CrIs may not have these properties for non-symmetric distributions (e.g., percentile-based CrIs). Figure 1 illustrates the potential difference between the HPD and percentile-based CrIs for a non-symmetric distribution.

Figure 1

Comparison of HPD and Percentile-Based CrIs



Additionally, to compare the performance of the Bayesian CrIs to frequentist CIs, bootstrap CIs were also considered. To summarize, the bootstrap for coefficient alpha can be outlined in three steps (Padilla, Divers, & Newton, 2012). Suppose that the data are $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right)$, where $\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ik},\right)$ is a set of variables. First, obtain the b^{th} bootstrap sample with replacement $\mathbf{X}^{(b)} = \left(\mathbf{x}_1^{(b)}, \mathbf{x}_2^{(b)}, \ldots, \mathbf{x}_n^{(b)}\right)$ from **X**. Second, compute and store the b^{th} estimate of coefficient alpha from $\mathbf{X}^{(b)}$ with

$$\alpha_c^{(b)} = \left(\frac{k}{k-1}\right)\left(1-\frac{\sum_{j=1}^{k}\sigma_{x_{jj}^{(b)}}}{\sigma_{X^{(b)}}^{2}}\right), \tag{82}$$

where $\sigma_{X^{(b)}}^{2}$ is the variance of the composite and $\sigma_{x_{jj}^{(b)}}$ is the variance of variable *j* for the b^{th} sample. Finally, compile the stored estimates $\alpha_c^{(1)}, \alpha_c^{(2)}, \ldots, \alpha_c^{(B)}$ to create the empirical sampling distribution (ESD) for the $b = 1, 2, \ldots, B$ bootstrap samples. The ESD can then be summarized to obtain statistical quantities for inference about α_c. The percentile bootstrap (PB) and the bias-corrected and accelerated (BCA) methods are commonly used in literature. The PB CI is estimated by obtaining the $\alpha / 2$ and $1-\alpha / 2$ percentiles from the ESD where α is the type I error probability. The BCA CI follows a similar process to the PB CI but the bounds are adjusted for bias and skew (i.e., acceleration) of the α_c ESD. See Efron and Tibshirani (1993) for details.

The focus here now is a simulation study to investigate the performance of the Bayesian CrIs and bootstrap CIs from above for the Bayesian and Bootstrap methods of estimating coefficient alpha. Specifically, interest is on the corresponding CrIs for the Bayesian coefficient alpha (Balpha 1) proposed by Padilla and Zhang (2011) and the new implementation of Bayesian coefficient alpha (Balpha 2) proposed here. The HPD CrI was not previously applied to Balpha 1 by Padilla & Zhang. As such, it will not be used as a point of comparison between Balpha 1 and 2. Instead, it will be tentatively explored with Balpha 2. Even so, the other two Balpha 1 and 2 CrIs will also be compared to the PB and BCA CIs.

METHODOLOGY

Simulation

A Monte Carlo simulation was used to investigate the Bayesian coefficient alpha. The simulation conditions were 5 (correlation type) × 4 (number of items) × 4 (number of item response categories) × 3 (distribution type) × 6 (sample size) for a total of 1,440 conditions. All simulated items were Likert-type (ordinal) or binary. For each simulation condition, 1,000 replications were obtained. The priors for the Bayesian coefficient alpha were set to be non-informative. This allows the Bayesian CrIs to be compared to frequentist coefficient alpha. The Bayesian CrIs and bootstrap CIs used 2,000 posterior draws and bootstrap samples, respectively.

An outline of the Monte Carlo simulation for the study is given below:

1. Select the structure of the $k \times k$ correlation matrix **P**, where *k* is the number of items.
2. Select a set of thresholds ***v*** to categorize items to a predetermined skewness and kurtosis.
3. Generate an $n \times k$ multivariate data matrix $\mathbf{Z} \sim N\left(0, \mathbf{P}\right)$, where *n* is the sample size.
4. Categorize the generated data **Z** using the thresholds in ***v*** to generate the dataset **X**. Each variable *x* in **X** is categorized by the thresholds as follows: $x = m$ if $v_m < z < v_{m+1}$ for $m = 0,\ 1,\ \ldots,\ M-1$ where $v_0 = -\infty$ and $v_M = \infty$, and *M* is the number of item response categories.
5. Compute the true population coefficient alpha $\left(\alpha_c\right)$ according to **P** and the thresholds in ***v***. See Maydeu-Olivares et. al. (2007) for details.
6. Estimate the Bayesian coefficient alpha CrIs from **X** as outlined above.
7. Determine if the CrIs contain the true population coefficient alpha $\left(\alpha_c\right)$.

Details of the simulation conditions are given below.

Conditions

Correlation Type (ρ)

Five different item correlation structures **P** were investigated. The first two correlation structures were from a parallel-item one factor model with common loadings $\lambda_1 = 0.55$ or $\lambda_2 = 0.705$. These two models will generate compound symmetric item correlation structures with $\rho = .30$ or .56, respectively. The next three correlation structures were generated from a congeneric item one-factor model with the following loadings: $\lambda_3 = 0.2,$ 0.3, 0.4, 0.5, 0.6; $\lambda_4 = 0.3,$ 0.4, 0.5, 0.6, 0.7; and $\lambda_5 = 0.4,$ 0.5, 0.6, 0.7, 0.8. These correlation structures were chosen for investigation as they were previously investigated by Padilla, Divers, & Newton (2012), based on work previously done by Maydeu-Olivares (2007), and allow for a greater range of conditions to explore the impact on coefficient alpha CI/CrIs. A summary for the population coefficient alphas these correlation structures produce is given in Table 2.

Table 2

Summary of Resultant Population Coefficient Alphas from Correlation Structure Types

Correlation Structure Type	Minimum α_c	Mean α_c	Maximum α_c
Parallel: λ = 0.55	.4827	.7513	.8870
Parallel: λ = 0.705	.7120	.8843	.9570
Congeneric: λ = 0.2 – 0.6	.2960	.5916	.7784
Congeneric: λ = 0.3 – 0.7	.4218	.7060	.8593
Congeneric: λ = 0.4 – 0.8	.5457	.7924	.9106

Number of Items (k)

Previous research on coefficient alpha has investigated numbers of items ranging from 3 to 40 (Barchard & Hakstian, 1997; Duhachek & Lacobucci, 2004; Maydeu-Olivares et al., 2007; Padilla et al., 2012; Padilla & Zhang, 2011; van Zyl et al., 2000). However, most of this previous research focused on numbers of items ranging from 5 to 20. Additionally, it is noted that going beyond 20 items for a measurement instrument reaches a point of diminishing returns with regards to coefficient alpha. To generalize to most of the previous research, the following number of items were investigated: $k = 5,\ 10,\ 15,\ 20$.

Number of Item Response Categories (IRC)

Previous research on coefficient alpha has investigated numbers of item response categories ranging from 2 to 7 (Maydeu-Olivares et al., 2007; Padilla et al., 2012; Romano et al., 2010; Yuan et al., 2003). To draw parity with previous research, the following common choices for response categories were investigated: $IRC = 2,\ 3,\ 5,\ 7$. For each response category, the first category was set to 0. For example, for an item with seven response categories, the categories are $m = 0, 1, 2, 3, 4, 5, 6$.

Distribution Type

Three different distribution types were investigated. When $IRC = 2$ (i.e., binary items), the thresholds for $\boldsymbol{v}$ were chosen such that the distributions had the following characteristics:

1. Type 1: $\text{skewness} = 0$ and $\text{kurtosis} = -2$
2. Type 2: $\text{skewness} = -1.70$ and $\text{kurtosis} = 0.88$
3. Type 3: $\text{skewness} = 0.41$ and $\text{kurtosis} = -1.83$

The Type 1 and 2 distributions for binary categorization were studied by Padilla et. al. (2012). The Type 3 distribution for binary categorization was studied by Maydeu-Olivares et. al. (2007).

When $IRC = 3, 5, 7$ (i.e., IRC > 2) the thresholds for $\boldsymbol{v}$ were be chosen so that the distributions had the following characteristics:

1. Type 1: $\text{skewness} = 0$ and $\text{kurtosis} = 0$
2. Type 2: $\text{skewness} = 0$ and $\text{kurtosis} = .88$
3. Type 3: $\text{skewness} = 0.97$ and $\text{kurtosis} = -0.20$

The Type 1 distribution for $IRC > 2$ categorizations was studied by Padilla et. al. (2012). The Type 2 and 3 distributions for $IRC > 2$ categorizations were studied by Maydeu-Olivares et. al. (2007). Investigating this breadth of distribution types allows for the assessment of how the Bayesian coefficient alpha responds to deviations from normality and draws parity with past research. The distribution types for items with two and five categories are presented in Figure 2.

Sample Size (n)

The following sample sizes were investigated: $n = 50, 100, 150, 200, 250, 300$. This selection is in line with most recent research on coefficient alpha (Duhachek & Lacobucci, 2004; Maydeu-Olivares et al., 2007; Padilla et al., 2012; Padilla & Zhang, 2011). The sample size selections above 200 are beyond the point of diminishing returns for coefficient alpha but were investigated as they are typically found in psychological/behavioral research.

Analysis

To evaluate the performance of the CIs/CrIs, coverage probability was used. Coverage probability is defined as the proportion of estimated CIs/CrIs that contain the true population coefficient alpha. Coverage probability was assessed using Bradley's (1978) liberal criterion defined by

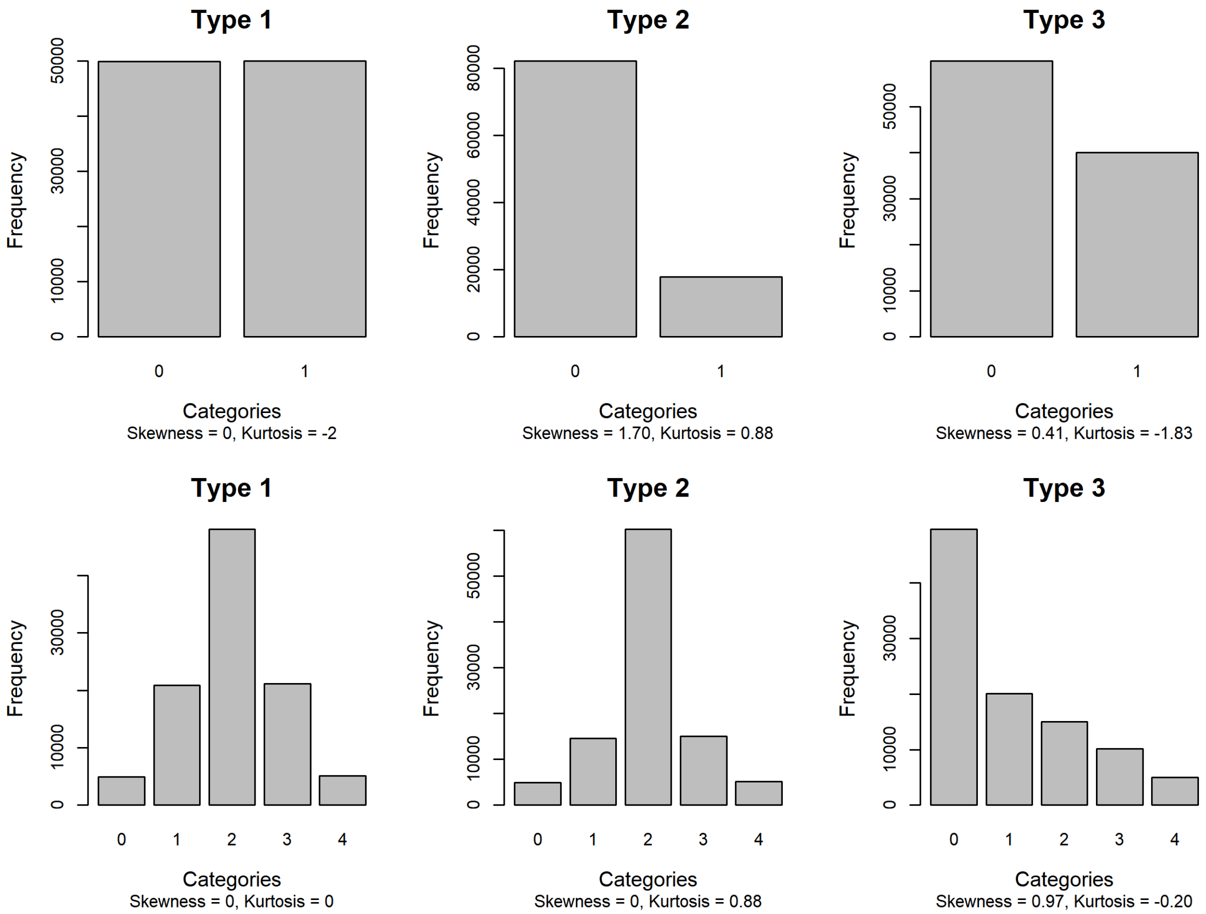
$$1 - 1.5\alpha \leq 1 - \alpha^* \leq 1 - 0.5\alpha, \qquad (83)$$

where $\alpha*$ is the true type I error probability. Here, the $100(1-\alpha)\%$ CIs/CrIs was estimated with $\alpha = .05$. Therefore, acceptable coverage probability is given by $[.925, .975]$.

For the benefit of prospective researchers, the implementation and interpretations of these coefficient alpha CIs/CrIs will be outlined with an example after the results of the simulation.

Figure 2

Bar Graphs of Dichotomous and 5-point Likert Items Utilized in Simulation



RESULTS

The simulated data was generated and analyzed using the *R* statistical package 4.2.2. The 5 (correlation type) × 4 (number of items) × 4 (number of item response categories) × 3 (distribution type) × 6 (sample size) simulation design resulted in 1,440 simulated conditions for coefficient alpha and its corresponding CIs/CrIs (i.e., Balpha 1 percentile, Balpha 1 NT, Balpha 2 percentile, Balpha 2 NT, balpha 2 HPD, PB, and BCa). There were 1,000 replications for each simulation condition. The Bayesian CrIs and bootstrap CIs used 2,000 posterior draws and bootstrap samples, respectively. Coverage probability performance for the CIs/CrIs was assessed using Bradley's (2015) liberal criterion (i.e., $\left[.925,.975\right]$). The forthcoming sections discuss the results for each simulation main effects and pairwise simulation conditions and are presented in figures $3-10$.

Coverage Probability

The results for the simulation main effects for all CI/CrI methods investigated are first discussed and presented in Figure 3. The results for the pairwise simulations conditions are then discussed individually for each CI/CrI method investigated and presented in Figures $4-10$.

Simulation Main Effects

Coverage for Correlation Type. In general, the CIs/CrIs tended to have acceptable coverage across the correlation types; see Figure 3. However, performance varied slightly among the CIs/CrIs in this condition. The BCa and PB CIs had stable performance across the correlation types, but the BCa tended to be closer to the target of .95. Even so, the BCa and PB CIs tended to be slightly below .95. In general, the Balpha 1 percentile and Balpha 1 NT CrIs tended to be slightly above .95. However, for the parallel model with a factor loading of 0.705, the Balpha 1 percentile CrI tended to be slightly below .95 whereas the Balpha 1 NT CrI tended to be on target. The Balpha 2 percentile, Balpha 2 NT, and Balpha 2 HPD CrIs performed similarly to one another with estimates near the target of .95 except in the parallel model case where the factor loading was 0.705. Here, each CrI tended to be slightly below .95, and the Balpha 2 HPD did not have acceptable coverage. Lastly, the bootstrap CIs were less variable than the Bayesian CrIs.

Coverage for Number of Items. In general, all CIs/CrIs tended to have acceptable coverage across the number of items. However, performance was not equal among the CIs/CrIs in this condition. The BCa CI had stable performance across all numbers of items. The PB CI tended to be slightly below .95 as number of items increased. Even so, the BCA and PB CIs tended to be slightly below .95. The Balpha 1 percentile and Balpha 1 NT CrIs also tended to be slightly above .95 as number of items increased. Here, the deviation from .95 was more than the PB CI. The Balpha 2 percentile, Balpha 2 NT, and Balpha 2 HPD CrIs tended to have better and more stable performance (i.e., closer to the target .95) as number of items increased. Lastly, the bootstrap CIs were less variable than the Bayesian CrIs.

Coverage for Number of Item Response Categories. In general, the CIs/CrIs tended to have acceptable coverage across the number of item response categories. However, performance was not equal among the CIs/CrIs in this condition. The BCa and PB CIs had stable performance across the number of item response categories, but the BCa tended to be closer to the .95 target. Even so, the BCA and PB CIs tended to be slightly below .95. All Bayesian CrIs tended to not have acceptable coverage when there were 2 item response categories. However, if there were at least 3 item response categories, all Bayesian CrIs tended to have acceptable and stable coverage. In these cases, the Balpha 1 percentile and Balpha 1 NT CrIs tended to be slightly above .95 and the Balpha 2 percentile, Balpha 2 NT, and Balpha 2 HPD CrIs tended to be slightly below .95.

Lastly, the bootstrap CIs and Bayesian CrIs had similar variability. The exception here is the Bayesian CrIs with 2 item responses categories had more variability than all other CIs/CrIs.

Coverage for Distribution Type. In general, the CIs/CrIs tended to have acceptable coverage across the distribution types. However, performance varied slightly among the CIs/CrIs in this condition. The BCa and PB CIs had stable performance across the number of item response categories, but the BCa tended to be closer to the target of .95. Even so, the BCA and PB CIs tended to be slightly below .95. The Balpha 1 percentile and Balpha 1 NT CrIs had performance similar to one another in that they tended to be slightly above .95 with distribution type 1 but were closer to the target with distribution types 2 and 3. The Balpha 2 percentile, Balpha 2 NT, and Balpha 2 HPD CrIs had performance similar to another. These tended to be slightly above .95 with distribution type 1 and slightly below .95 with distribution types 2 and 3. There was a larger amount of variability with distribution type 2 and the Balpha 2 HPD CrI did not have acceptable coverage in this case. Lastly, the bootstrap CIs and Bayesian CrIs had similar variability. The exception here is the Bayesian CrIs with distribution type 2 had more variability than all other CIs/CrIs.

Coverage for Sample Size. In general, the CIs/CrIs tended to have acceptable coverage across the sample sizes. However, performance was not equal among the CIs/CrIs in this condition. The BCa and PB CIs tended to be slightly below .95 and did not have acceptable coverage when sample size was 50. However, they had acceptable coverage when sample size was at least 100. The Balpha 1 percentile and Balpha 1 NT CrIs tended to be slightly above .95 and did not have acceptable coverage when sample size was 50. However, they had acceptable coverage when sample size was at least 100. The Balpha 2 percentile, Balpha 2 NT, and Balpha

2 HPD CrIs had acceptable coverage across sample sizes but tended to be slightly below .95 as sample size increased. Lastly, the bootstrap CIs were less variable than the Bayesian CrIs.

Pairwise Simulation Effects for Each CI

PB CI. Figure 4 shows the coverage probability performance for the PB CI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were some instances of unacceptable performance that stand out. Most noticeably, simulation condition combinations that involved sample size had unacceptable coverage when sample size was 50. With greater sample sizes this method tends to stabilize around the target coverage of .95 for all simulation condition combinations. Additionally, there was some instances of unacceptable coverage with distribution type 2. These occurred when distribution type 2 was with 2 item response categories or a congenic model with factor loadings of $0.3-0.7$.

BCa CI. Figure 5 shows the coverage probability performance for the BCa CI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were some instances of unacceptable performance that stand out. Most noticeably, many simulation condition combinations that involved sample size had unacceptable coverage when sample size was 50. The exceptions where with a parallel model with a factor loading of 0.705 or with distribution types 1. With greater sample sizes this method tends to stabilize around the target coverage of .95 for all simulation condition combinations; with less variability than the PB CI.

Balpha 1 Percentile CrI. Figure 6 shows the coverage probability performance for the Balpha 1 percentile CrI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were two main simulation conditions that had instances of unacceptable performance that stand out. First, many simulation condition combinations that involved 2 item response categories had unacceptable coverage. Exceptions to this was when there was a congeneric model with a factor loading of $0.2-0.6$ or distribution types 1 and 3. Second, a sample size of 50 had several simulation condition combinations that had unacceptable coverage. Specifically, for a sample size of 50, there was unacceptable coverage when there were 20 items, for distribution type 1, for congeneric models that had factor loadings $0.2-0.6$ and $0.3-0.7$, and a parallel model that had a factor loading of 0.55. Additionally, there was unacceptable coverage for a parallel model with a factor loading of 0.705 with distribution types 1 and 2.

Balpha 1 NT CrI. Figure 7 shows the coverage probability performance for the Balpha 1 NT CrI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were two main simulation conditions that had instances of unacceptable performance that stand out. First, many simulation conditions that involved 2 item response categories had unacceptable coverage. Exceptions to this were when there were 15 items, sample size was 100, when there was a there was a congeneric model with a factor loading $0.2-0.6$, or distribution types 1 and 3. Second, a sample size of 50 had several combinations of simulation conditions with unacceptable coverage. Specifically, when there was more than 10 items, for all levels of correlation type, or for all levels of distribution type. Additionally, there was unacceptable coverage when there were 20 items combined with a congeneric model with a factor loading of $0.3-0.7$.

Balpha 2 Percentile CrI. Figure 8 shows the coverage probability performance for the Balpha 2 percentile CrI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were two main simulation conditions that had instances of unacceptable performance that stand out. First, many simulation conditions that involved 2 item response categories had unacceptable coverage. Exceptions to this were a congeneric model with a factor loading of $0.2-0.6$, distribution types 1 and 3, or sample size of 50. Second, there were also some instances of unacceptable coverage for simulation condition combinations that involved correlation type when there was a parallel model with a factor loading of 0.705. This includes instances when there were less than 15 items, distribution types 2 and 3, or when sample size was 250. Additionally, there were some instances of unacceptable coverage for distribution type 2 when sample size was 100, 250 or 300.

Balpha 2 NT CrI. Figure 9 shows the coverage probability performance for the Balpha 2 NT CrI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were two main simulation conditions that had instances of unacceptable performance that stand out. First, many simulation condition combinations that involved 2 item response categories had unacceptable coverage. Exceptions to this were when there was a congeneric model with a factor loading of $0.2-0.6$ or distribution types 1 and 3. Second, distribution type also had several combinations of simulation conditions with unacceptable coverage. For distribution type 2, there was unacceptable coverage when there were 10 items, a congeneric model with factor loadings $0.4-0.8$, a parallel model with a factor loading of 0.705, or sample sizes 100 and 300. For distribution type 3, there was unacceptable coverage for a parallel model with a factor loading of 0.705. Additionally, a parallel model with a factor loading of 0.705 did not have acceptable coverage when there was 10 items or sample size was greater than 200.

Balpha 2 HPD CrI. Figure 10 shows the coverage probability performance for the Balpha 2 HPD CrI for all pairwise simulation conditions. In general, this method had acceptable coverage across the simulation condition combinations but there were three main simulation conditions that had instances of unacceptable performance that stand out. First, many simulation condition combinations that involved 2 item response categories had unacceptable coverage. Exceptions to this occurred with distribution types 1 and 3. Second, distribution type also had several simulation conditions combinations with unacceptable coverage. For distribution type 2, there was unacceptable coverage if there were more than 5 items, a congeneric model with factor loading $0.4-0.8$, parallel models with factor loadings of 0.55 and 0.705, or sample sizes larger than 50. Additionally, for distribution type 3, there was unacceptable coverage for a parallel model with factor loading of 0.705. Lastly, the parallel model with a factor loading of 0.705 also did not have acceptable coverage when there were 20 items, there were 5-item response categories, or sample size was greater than 100.

Summary

From the bootstrap methods, the BCa and PB CIs were generally robust to the conditions of the simulation except for conditions that involved a sample size of 50. In such cases, these CIs tended to have unacceptable coverage that trended slightly below .95. However, these CIs tended to recover and approach the target when sample size was at least 100. The BCa CI tended to outperform the PB CI by being closer to the target and by having less variability to certain conditions. Notably, the PB CI was observed to have a sensitivity to distribution type 2 by not having acceptable coverage when there were 2 item response categories.

The Balpha 1 percentile and Balpha 1 NT CrIs were generally robust to the conditions of the simulation with certain exceptions. Notably, these CrIs had a sensitivity to conditions that involved having 2 item response categories and tended to not have acceptable coverage and/or more variability in such cases. Additionally, these CrIs were sensitive to conditions that involved a sample size of 50. At a sample size of 50 and across all the number of item response categories, the Balpha 1 percentile and Balpa 1 NT CrIs tended to have unacceptable coverage that trended slightly above .95. The effect a sample size of 50 had was also seen across all distribution types, with distribution type 2 showing the most variability. The sensitivity to distribution type 2 for these CrIs can also be seen by some lack of acceptable coverage with condition combinations with the parallel model case where the factor loading was 0.705 and cases with 2 item response categories.

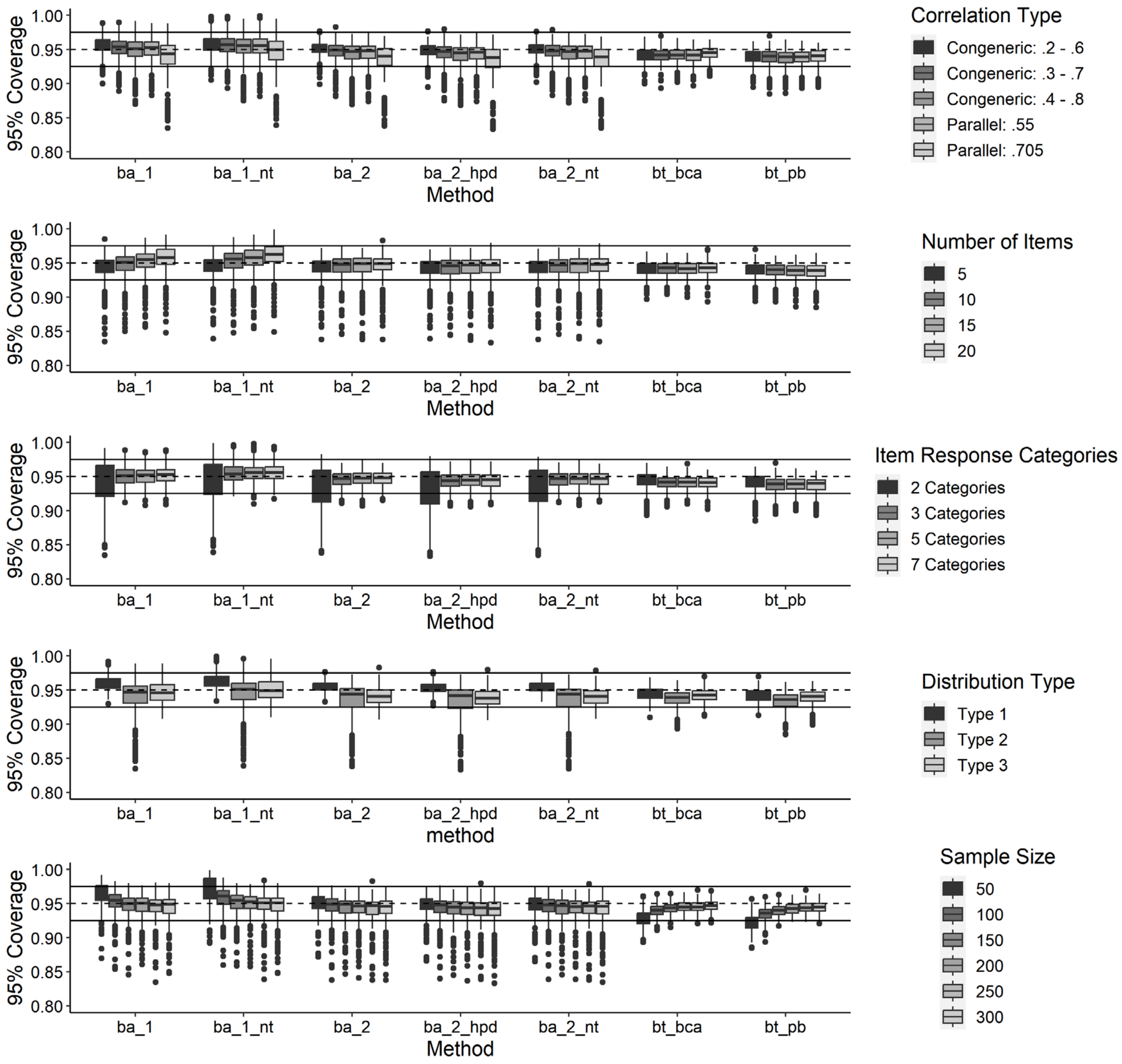
Like the Balpha 1 percentile and Balpha 1 NT CrIs, the Balpha 2 percentile, Balpha 2 NT, and Balpha 2 HPD CrIs were generally robust to the conditions of the simulation with certain exceptions. Additionally, like the Balpha 1 CrIs, the Balpha 2 CrIs were also sensitive to conditions that involved having 2 item response categories by tending to not have acceptable coverage and/or more variability in such cases. The Balpha 2 CrIs do not share a sensitivity to sample size but were generally more sensitive to condition combinations that involve a parallel model where the factor loading was 0.705. This sensitivity resulted in a lack of acceptable coverage and/or more variability. There also appears to be a greater sensitivity to distribution type 2 that also resulted in a lack of acceptable coverage and/or greater variability when combined with number of items. The Balpha 2 percentile and Balpha 2 NT CrIs tended to perform similarly to each other, but the Balpha 2 HPD CrI tended to have less instances of acceptable coverage.

To further summarize, all CIs/CrIs investigated tended to have acceptable coverage to the simulation conditions. However, there are certain nuances to each method. The bootstrap CIs are most sensitive to conditions with a sample size of 50. The Balpha 1 and Balpha 1 NT CrIs are most sensitive to conditions with 2 item response categories and/or sample size of 50. The

Balpha 2, Balpha 2 NT, and Balpha 2 HPD CrIs are most sensitive to conditions with 2 item response categories, a parallel model with a factor loading of 0.705, and/or distribution type 2.

Figure 3

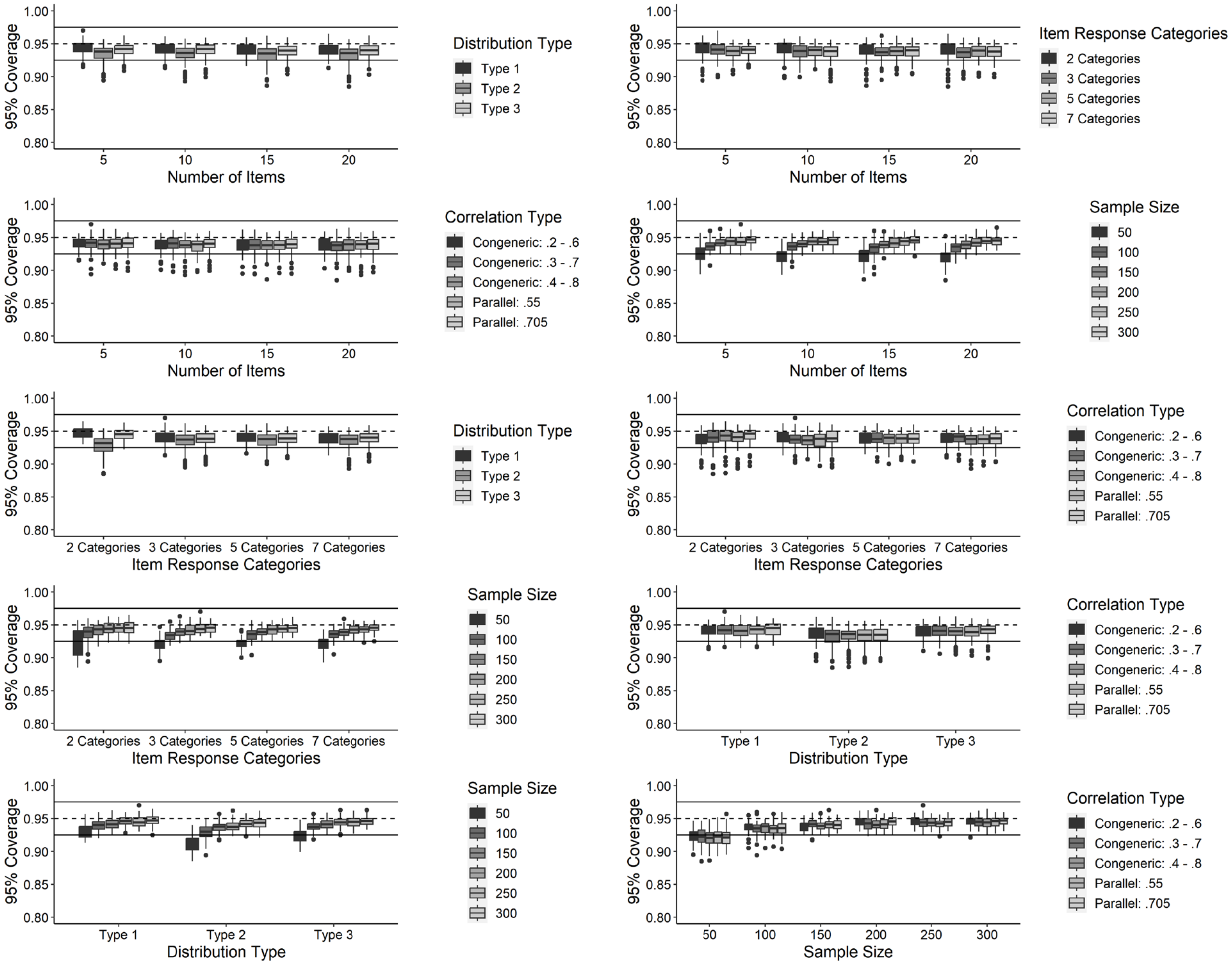
Distribution of the 95% Confidence/Credible Intervals for All Simulation Main Effects



Note. Balpha 1 percentile (ba_1), balpha 1 normal theory (ba_1_nt), balpha 2 percentile (ba_2), balpha 2 highest probability density (ba_2_hpd), balpha 2 normal theory (ba_2_nt), bias-corrected and accelerated (bt_bca), percentile bootstrap (bt_pb); Bootstrap methods (bt_bca and bt_pb) based on 2,000 bootstrap samples; Bayesian methods based on 2,000 posterior draws; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 4

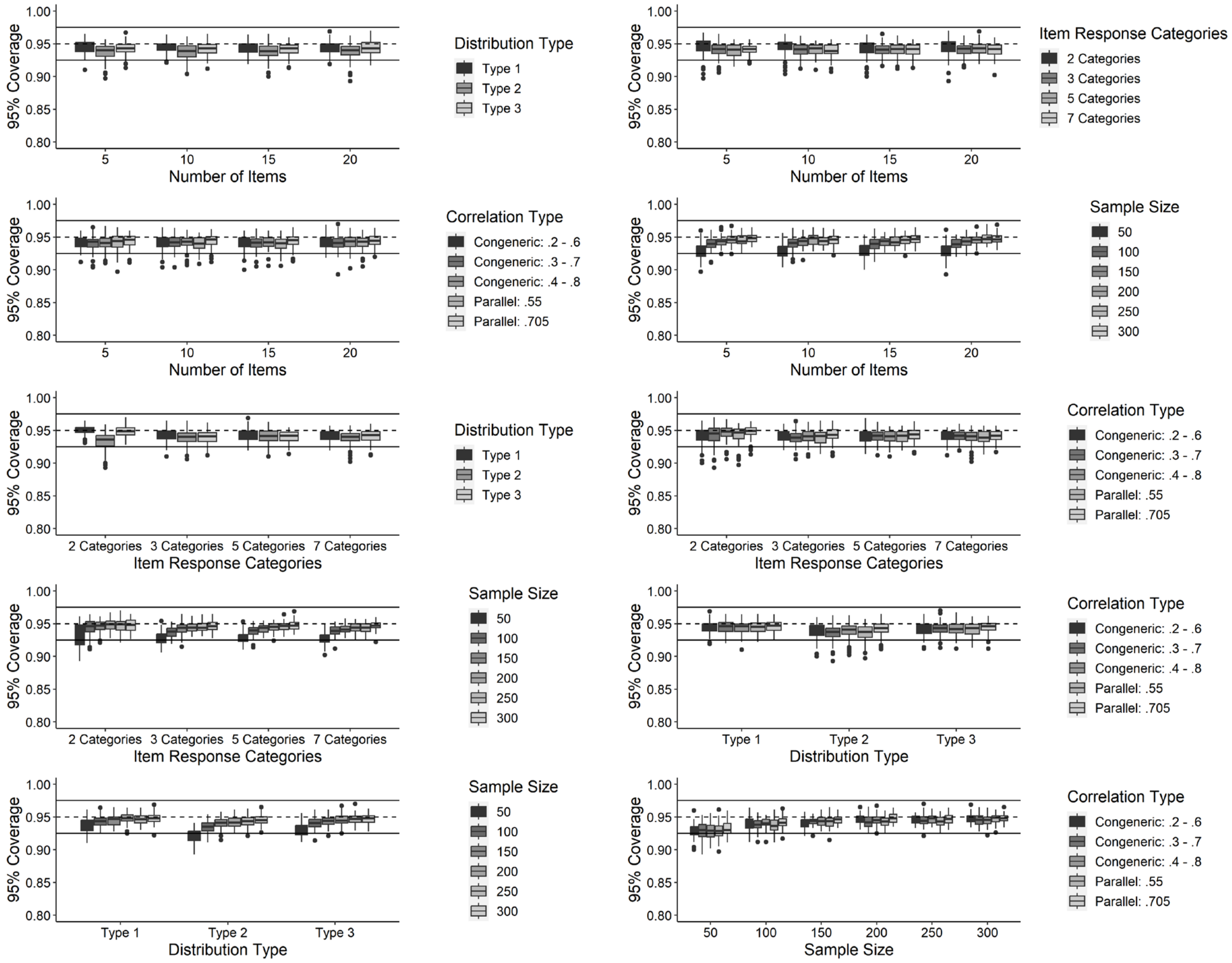
Percentile Bootstrap 95% Confidence Interval Coverage for all Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Percentile bootstrap confidence interval based on 2,000 bootstrap samples; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 5

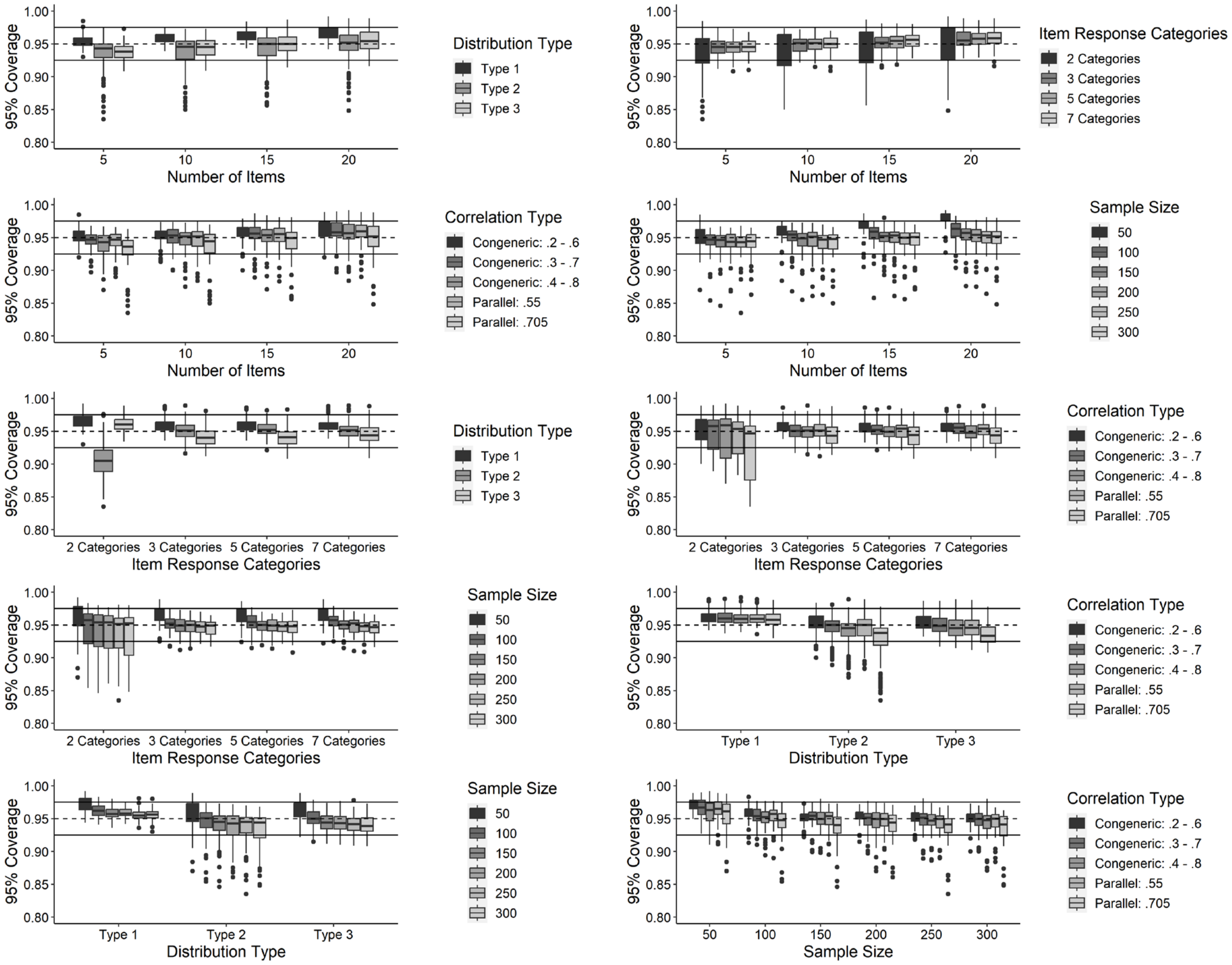
Bias-Corrected and Accelerated 95% Confidence Interval Coverage for All Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Bias-corrected and accelerated confidence interval based on 2,000 bootstrap samples; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 6

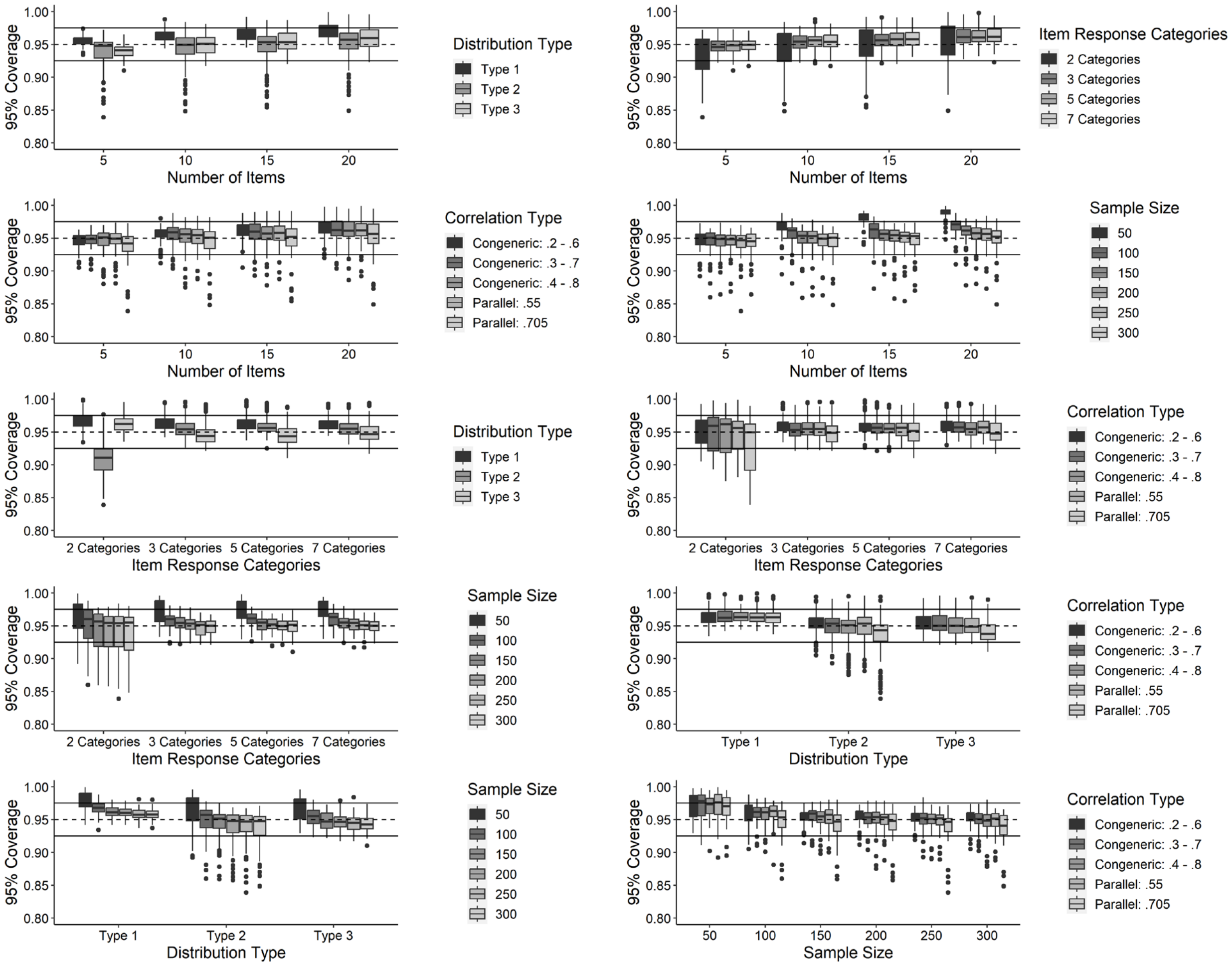
Balpha 1 Percentile 95% Credible Interval Coverage for All Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Balpha 1 percentile credible interval based on 2,000 posterior draws; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 7

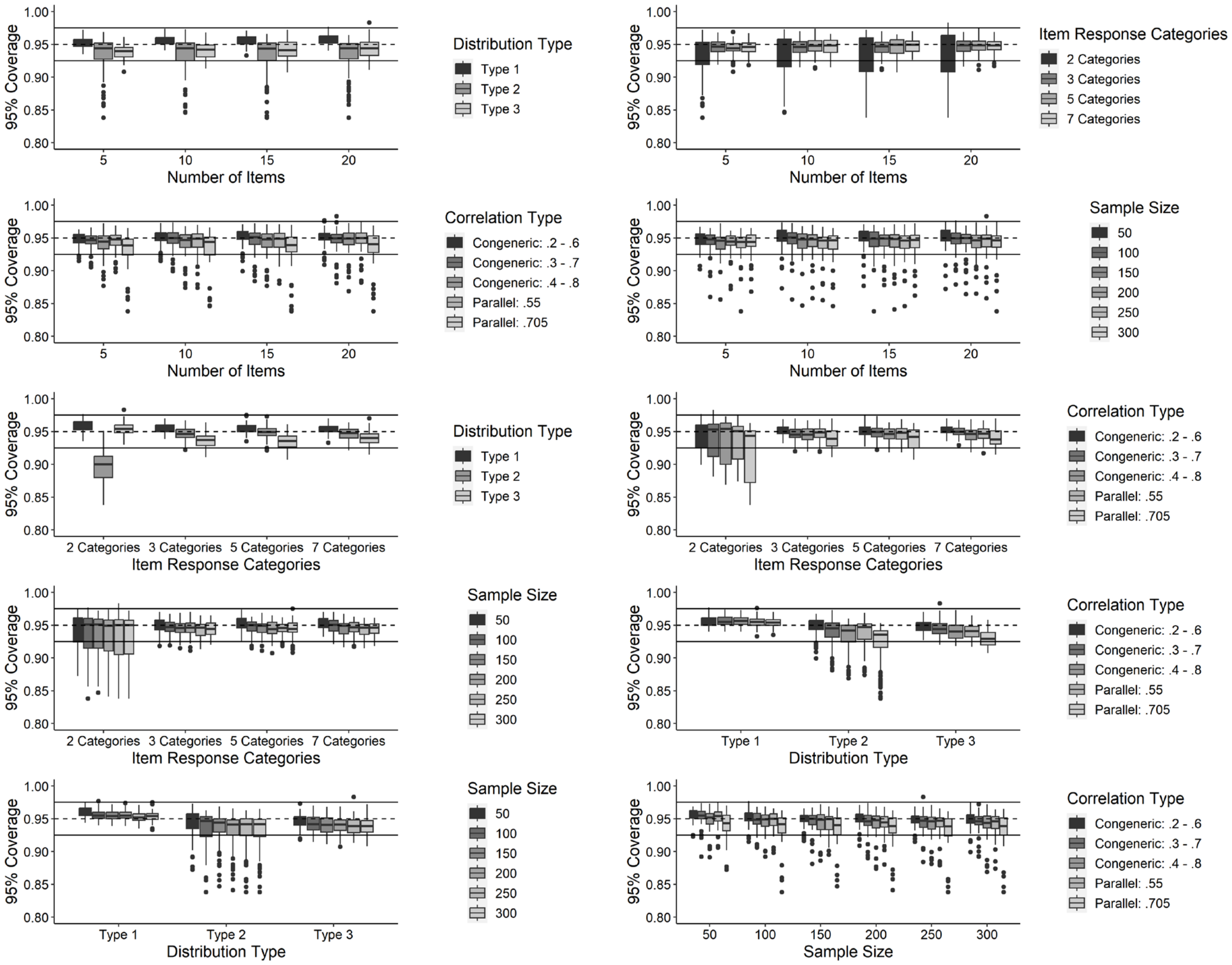
Balpha 1 Normal Theory 95% Credible Interval Coverage for All Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Balpha 1 normal theory credible interval based on 2,000 posterior draws; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 8

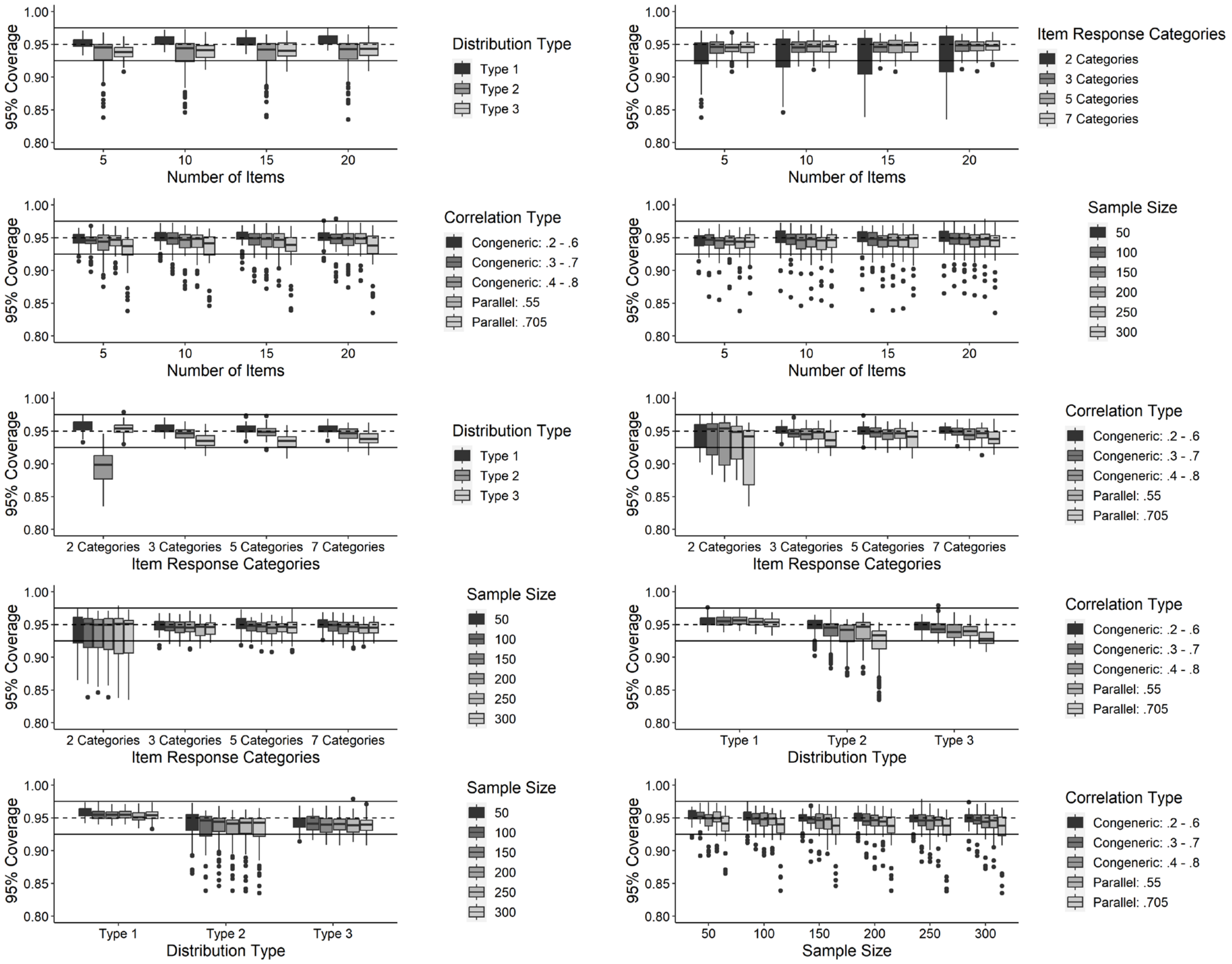
Balpha 2 Percentile 95% Credible Interval Coverage for all Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Balpha 2 percentile credible interval based on 2,000 posterior draws; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 9

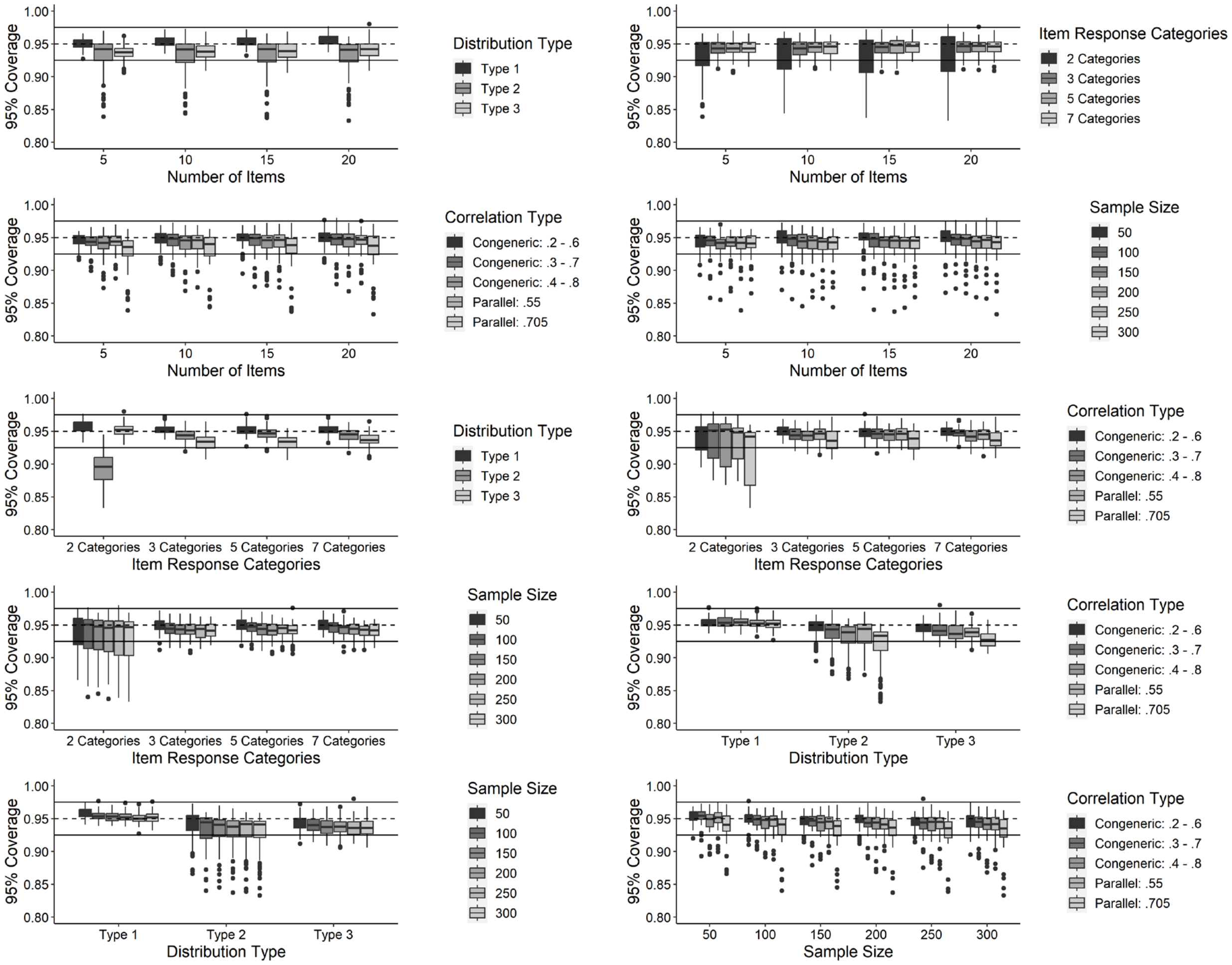
Balpha 2 Normal Theory 95% Credible Interval Coverage for All Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Balpha 2 normal theory credible interval based on 2,000 posterior draws; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Figure 10

Balpha 2 Highest Probability Density 95% Credible Interval Coverage for All Pairwise Simulation Conditions



Note. Distributions with 2 categories: type 1 (skewness = 1, kurtosis = -2), type 2 (skewness = -1.70, kurtosis = 0.88), type 3 (skewness = 0.41, kurtosis = -1.83); Distributions with > 2 categories: type 1 (skewness = 0, kurtosis = 0), type 2 (skewness = 0, kurtosis = 0.88), type 3 (skewness = 0.97, kurtosis = -0.20); Balpha 2 highest probability density credible interval based on 2,000 posterior draws; Dashed line at .95 and solid lines at acceptable coverage of [.925, .975].

Application Example

The coefficient alpha CIs/CrIs investigated in the simulation were also applied to unpublished real data. This data consists of responses to the Short Form Big Five Inventory-2 (BFI-2-S), which consists of 30 5-point Likert-scale items and was completed by 222 respondents. The BFI-2-S is a self-report measurement instruments used to assess a respondent's measurement on the constructs of 1) agreeableness, 2) open mindedness, 3) conscientiousness, 4) negative emotionality, and 5) extraversion (Soto & John, 2017).

For demonstration purposes, the coefficient alpha CIs/CrIs were applied to the extraversion sub-measure of the BFI-2-S. This sub-measure consists of items 1, 6, 11, 16, 21, and 26 from the BFI-2-S. A CFA was performed to investigate the unidimensionality of extraversion. Due to convergence issues, item 16 was dropped from the CFA and the subsequent coefficient alpha CI/CrI estimations. The results of this CFA can be found in Tables 3 and 4. The bootstrap CIs involved generating the bootstrap distribution (ESD) with corresponding CIs and were based on 2,000 bootstrap samples. The Bayesian CrIs involved generating the posterior distribution with corresponding CrIs and were based on 2,000 posterior draws with a non-informative prior. The results of the 95% coefficient alpha CI/CrIs are summarized in Table 5 and described below.

For the extraversion sub-measure, the estimated coefficient alpha was .52, the average bootstrap coefficient alpha was .51, the average Balpha 1 estimate was .52, and the average Balpha 2 estimate was .52. In terms of interval width, the bootstrap methods had smaller widths compared to the Bayesian methods; with the PB having the smallest width (at four decimal places). Among the Bayesian methods, the Balpha 2 methods had smaller widths than the Balpha 1 methods; with the HPD having the smallest width (at three decimal places).

It is important to note the differences in interpretation between the bootstrap CIs and Bayesian CrIs. For example, using the PB method, the interpretation is that for all CIs created in the same way, we can be 95% confident that the true parameter of coefficient alpha for the measurement instrument is within $[0.42, 0.60]$. In contrast, using the Balpha 2 percentile method, the interpretation is that there is 95% probability that the true parameter of coefficient alpha for the measurement instrument is within $[0.42, 0.62]$.

Table 3

CFA Standardized Loadings for Extraversion of the Short Form Big Five Inventory-2

Extraversion	
Item	Estimate
1	0.38**
6	0.24**
11	0.20**
21	0.55**
26	0.67**

Note. Model: χ^2 (5, *N*=222) = 96.98. Null: χ^2 (10, *N*=222) = 159.65. ** indicates $p < .05$. ML estimation. Item 16 not included due to convergence issues.

Table 4

Model Fit Indices of CFA for Extraversion of the Short Form Big Five Inventory-2

χ^2	df	*p*	CFI	TLI	SRMR	RMSEA
96.98	5	0.00	**0.39**	**-0.23**	**0.14**	**0.29**

Note. Null Model: χ^2 (10, *N*=222) = 159.65. ML estimation. Values in bold did not meet the liberal criteria.

Table 5

95% Coefficient Alpha Confidence/Credible Intervals for the Extraversion Sub-Measure of the Short Form Big Five Inventory-2

Method	Estimate	Lower	Upper	Width
Percentile Bootstrap	.52	.42	.60	.18
Bias-Corrected Accelerated	.52	.42	.60	.18
Balpha 1 Percentile	.52	.41	.61	.20
Balpha 1 Normal Theory	.52	.41	.62	.21
Balpha 2 Percentile	.52	.42	.62	.20
Balpha 2 Normal Theory	.52	.42	.62	.20
Balpha 2 High Probability Density	.52	.42	.62	.20

Note. Bootstrap methods (Percentile Bootstrap and Bias-Corrected Accelerated) based on 2,000 bootstrap samples; Bayesian methods based on 2,000 posterior draws.

DISCUSSION

Measurement is a fundamental component of the behavioral/social sciences that has plenty of nuance making it difficult to implement. This difficulty stems from the behavioral/social sciences typically focusing on constructs that must be indirectly measured. Measuring these constructs is usually done through a measurement instrument that records behaviors and/or responses that are manifestations of the construct of interest. The indirect nature of these measurement instruments leads to measurement error, which is reflected in the instrument psychometric properties of reliability and validity. While both psychometric properties are important to a measurement instrument, the focus of this paper is reliability. The aspects and current development of reliability were outlined in this paper with a focus on the most popular form of reliability, coefficient alpha.

Coefficient alpha has seen the most development as a reliability statistic due to its popularity. Coefficient alpha's popularity stems from its computational ease as it requires only one measurement instrument administration and the instrument's item covariance matrix. However, it should be noted that coefficient alpha is only equal to reliability if the items of the measurement instrument are at least (essentially) tau-equivalent (i.e., have a compound symmetric item covariance matrix). If this is not the case, coefficient alpha is a lower bound to reliability. Even so, coefficient alpha has developed to the point where its distribution is understood enough to allow for the estimation of frequentist CIs and Bayesian CrIs.

An early Bayesian coefficient alpha (Balpha 1) was proposed by Padilla and Zhang (2011). Balpha 1 indirectly generates the posterior for coefficient alpha via the posterior for the item covariance matrix. Therefore, Balpha 1 uses a prior for the item covariance matrix. If a prior is not needed and/or available, then a non-informative prior is easy to obtain for the item covariance matrix. Unfortunately, the item covariance matrix for a measurement instrument is typically not reported in the literature, so Balpha 1 is difficult to implement in practice if a prior is needed. To reiterate, Balpha 1 requires a prior for the item covariance matrix as opposed to a prior for coefficient alpha.

Building on Balpha 1, a new derivation of Bayesian coefficient alpha (Balpha 2) was proposed in this paper. Unlike Balpha 1, Balpha 2 is based on coefficient alpha having a posterior normal distribution based on a posterior mean $\left(\alpha_c\right)$ and variance $\left(\sigma_\alpha^2\right)$ with each having a corresponding prior. Coefficient alpha is readily reported in the literature, so using a reported coefficient alpha as a prior is straightforward. However, variance for coefficient alpha is not reported in the literature. Fortunately, item variances are usually reported and a method for using item variances was presented. The method essentially averages the item variances to a common variance, which can then be used to get a prior for the coefficient alpha variance. From here, interest is in how Balpha 2 performs.

A Monte Carlo simulation was used to assess the performance of Balpha 2 compared to Balpha 1 and bootstrap coefficient alpha via CIs/CrIs. Data were generated with the following conditions: correlation type, number of items (*k*), number of item response categories (*IRC*), distribution type, and sample size (*n*). The bootstrap CIs were estimated with PB and BCA methods, the Balpha 1 CrIs were estimated with percentile and NT methods, and the Balpha 2 CrIs were estimated using percentile, NT, and HPD methods. Coverage probability performance was assessed for each CI/CrI.

The bootstrap CIs (PB and BCA) were generally robust to conditions of the simulation but were negatively impacted by certain conditions. Most notably, the bootstrap CIs tended to have unacceptable coverage probability in conditions with the smallest sample size investigated

($n = 50$). However, as sample size increased, the bootstrap CIs tended to stabilize toward acceptable coverage. The PB CI was noted to not have acceptable coverage when there were 2 item response categories with distribution type 2. The type 2 distribution with 2 item response categories had the highest skew where one response category had a much lower response. Although the BCA CI was also negatively impacted by the type 2 distribution with 2 item response categories, it still maintained acceptable coverage. If there were more than 2 item response categories ($IRC > 2$) with distribution type 2, both the PB and BCA CIs had adequate coverage.

The Balpha 1 CrIs (percentile and NT) were generally robust to the conditions of the simulation but were negatively impacted by certain conditions. Most notably, the Balpha 1 CrIs tended to have unacceptable coverage probability in conditions that involved having 2 item response categories. There were also instances of unacceptable coverage in condition combinations with the smallest sample size investigated ($n = 50$), more than 10 items $(k > 10)$, and distribution type 2. Although these conditions affected both the Balpha 1 percentile and Balpha 1 NT methods, the Balpha 1 NT tended to have better coverage than the Balpha 1 percentile.

The Balpha 2 CrIs (percentile, NT, and HPD) were generally robust to the conditions of the simulation but were negatively impacted by certain conditions. As with balpha 1, balpha 2 also tended to have unacceptable coverage probability in conditions that involved having 2 item response categories. There were also instances of unacceptable coverage with conditions using a parallel model with a factor loading of 0.705 or distribution type 2. The difference in sensitivities to certain conditions between Balpha 1 and Balpha 2 may be due to the number parameters used to generate the posterior for each method. Balpha 1 indirectly generates the posterior for coefficient alpha via the posterior for the item covariance matrix. Balpha 2 directly generates the coefficient alpha posterior with a posterior mean and variance as reflected by the posterior coefficient alpha $\left(\alpha_c\right)$ and corresponding variance $\left(\sigma_\alpha^2\right)$, respectively. It appears that directly generating the posterior distribution with a mean to center it and a variance to stabilize it gives Balpha 2 an advantage. This is reflected in that the Balpha 2 CrIs tended to behave better than their Balpha 1 equivalents with coverage probability that tended to be closer and tighter around the target of .95. These conditions affect all the Balpha 2 CrIs. However, the Balpha 2 percentile and NT perform similarly to one another and had better coverage than the Balpha 2 HPD.

The CI/CrIs were impacted by different conditions. The bootstrap CIs were impacted primarily by sample size as they required a sample size of 100 or more to have acceptable coverage probability. This is not surprising as the bootstrap is a large sample nonparametric method, and estimation of a correlation matrix requires a healthy sample size (Raykov, 1998).

On the other hand, the Bayesian CrIs were mainly impacted by item response categories, distribution type, and correlation type. Specifically, the CrIs were impacted by items with 2 items response categories, came from distribution type 2, and came from a parallel model with a factor loading of 0.705. A common feature of these conditions is that they have some form of range restriction, and range restrictions attenuate variability. First, the Balpha CrIs are based on making direct draws from a normal posterior distribution. However, items with 2 item response categories cannot achieve normality and are at odds with the draws from the normal posterior distribution. This is because items with 2 item response categories effectively have range restrictions that cannot accurately reflect the variability in the normal distribution. Second, items with skewed distributions tend to have range restrictions because of floor/ceiling effects. This can clearly be seen for the items with skewness greater or equal to 0.97 in Figure 2. Lastly, a parallel model with a factor loading of 0.705 had a corresponding compound symmetric item correlation structure with $\rho = .56$ that created the strongest coefficient alphas with an average coefficient alpha of 0.88 (see table 2). This created a range restriction at the upper limit of 1 for coefficient alpha. Any one of these conditions negatively impacted the CrIs, but any combination of them exacerbated the impact. Again, range restrictions attenuate variance and accurate variance is required to have accurate CrIs.

For applications purposes, the BCA CI and Balpha 2 percentile/NT are recommended depending on the conditions at play for coefficient alpha. The BCA CI is recommended if the sample size is large enough $\left(n \geq 100\right)$ and a Bayesian prior is not necessary and/or available. On the other hand, the Balpha 2 percentile/NT CrIs are suggested if the sample size is small $\left(n < 100\right)$, there are more than 2 item response categories, and a Bayesian prior is not necessary and/or available. Note that in situations where a prior is not necessary and/or available, Bayesian CrIs with non-informative priors achieve the same conclusions as their frequentist counterparts. Additionally, Bayesian CrIs always maintain their interpretation regardless of having informative or non-informative priors. However, if a Bayesian prior is necessary and/or available, then the Balpha 2 percentile/NT CrIs are recommended. It should be noted that the recommendations presented here are based and limited to the conditions investigated in the simulation. Even so, the breadth of conditions investigated cover a widespread number of general situations relevant to applied settings.

There are several possibilities to expand on the findings presented in this paper. The HPD method for interval estimation was introduced and applied to Balpha 2. Further insight into the performance of the HPD for other estimation methods of coefficient alpha was not investigated here as the focus of the paper was the development Balpha 2. Although the HPD did not perform as well as the Balpha 2 percentile and NT CrIs, it would be pertinent to explore how the HPD performs for Balpha 1 or the bootstrap coefficient alpha. For further parity among the estimation methods, it would be worth exploring how a NT bootstrap CI of coefficient alpha performs compared to these other methods. Additionally, given that some CI/CrIs methods were impacted by sample size, it would be of interest to explore sample size sensitivity. For the Bootstrap CIs, this means exploring at what sample size acceptable coverage is achieved and at what sample size acceptable coverage is maximized or reaches a point of diminishing returns. For the CrIs, this means exploring how small the sample size can be and still maintain acceptable coverage. The CrIs were impacted by items with 2 item response categories. In this situation, it may be fruitful to explore a Gibbs sampler when making posterior draws (Hoff, 2009). Unlike the MCMC method used here that directly makes draws for a normal posterior, an advantage of the Gibbs sampler is that it cycles through the data at each draw (i.e., iteration). This cycling through the data allows the data to have more of an impact on the posterior, which in turn can help improve the situation with items with 2 response categories. Incidentally, cycling through the data at each draw may also help with skewed items (i.e., non-normality). Finally, another suggestion would be to explore the performance of the CIs/CrIs for the difference between coefficient alphas (i.e., $\alpha_{c1} - \alpha_{c2}$). This is because there may be situations in which researchers would like to compare coefficient alphas between groups like males vs. females, treatment vs. no treatment, etc.

REFERENCES

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*(1), 32-50. doi: 10.1037/0003-066X.63.1.32

APA. (2020a). Graduate study in psychology. Retrieved from https://gradstudy.apa.org/

APA. (2020b). Quantitative and qualitative programs in the United States of America or Canada. Retrieved from https://www.apadivisions.org/division-5/resources/doctoral

Barchard, K. A., & Hakstian, A. R. (1997). The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivariate Behavioral Research, 32*(2), 169-191. doi:10.1207/s15327906mbr3202_4

Beaulieu-Prevost, D. (2006). Confidence intervals: From tests of statistical significance to confidence itnervals, range hypotheses, and substantial effects. *Tutorials in Quantitativve Methods for Psychology, 2*(1), 11-19. doi:10.20982/tqmp.02.1.p011

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behavior*. doi:10.1038/s41562-017-0189-z

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi:10.1111/j.2044-8317.1978.tb00581.x

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods, 19*(4), 651-682. doi:10.1177/1094428116656239

Cohen, R. J. (2010). *Psychological testing and assessment: An introduction to tests and measurement* (7th ed.). Boston: McGraw-Hill Higher Education.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The Hopkins symptom checklist (HSCL): A self-report symptom inventory. *Behavioral Science, 19*(1), 1-15. doi:10.1002/bs.3830190102

DeVellis, R. F. (2017). *Scale development : theory and applications* (Fourth edition. ed.). Los Angeles: SAGE.

Duhachek, A., & Lacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *The Journal of Applied Psychology, 89*(5). doi:10.1037/0021-9010.89.5.792

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*(3), 357-370. doi:10.1007/BF02289499

Fisher, R. A. (1950). *Statistical methods for research workers* (11th ed.). Edinburgh,: Oliver and Boyd.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics : an introduction* (Second edition. ed.). Los Angeles: SAGE.

Graham, M. J. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66*(6), 930-944. doi:10.1177/0013164406288165

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255–282 doi:https://doi.org/10.1007/BF02288892

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231. doi:10.1007/BF02291840

Hoff, P. D. (2009). A first course in Bayesian statistical methods. London ; New York: Springer.

Hogg, R. V., Tanis, E. A., & Zimmerman, D. L. (2015). *Probability and statistical inference* (Ninth edition. ed.). Boston: Pearson.

Koning, A. J., & Franses, P. H. (2003). *Confidence intervals for comparing Cronbach's coefficient alpha values*. Erasmus University and the Erasmus School of Economics (ESE) at Erasmus University Rotterdam.: Erasmus Research Institute of Management.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160. doi:https://doi.org/10.1007/BF02288391

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

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

McDonald, R. P. (1970). The theoretical foundations of principal components factor analysis, canoncical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*(1), 1-21. doi:https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

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

Padilla, M. A., & Zhang, G. (2011). Estimating internal consistency using Bayesian methods. *Journal of Modern Applied Statistical Methods, 10*(1), 277-286. doi:10.22237/jmasm/1304223840

Raykov, T. (1997). Estimation of composite for congeneric measures. *Applied Psychological Measurement, 21*(2), 173-184. doi:10.1177/01466216970212006

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. doi:10.1177/014662169802200406

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*: Routledge.

Revelle, W. (2022). Psych: procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois. R package version 2.2.5, https://CRAN.R-project.org/package=psych.

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. doi:10.1177/0013164409355690

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99-103.

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

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69-81. doi:10.1016/j.jrp.2017.02.004

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Education, 2*, 53-55. doi:10.5116/ijme.4dfb.8dfd

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

Walker, D. A. (2006). A comparison of the Spearman-Brown and Flanagan-Rulon formulas for split half reliability under various variance parameter conditions. *Journal of Modern Applied Statistical Methods, 5*(2), 443-451. doi:10.22237/jmasm/1162354620

Wilkenson, L. (1999). Statistical methods in psychology journals. *American Psychologist, 54*(8), 594-604. doi:10.1037/0003-066X.54.8.594

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42*(4), 343–357. doi:https://doi.org/10.1007/BF02295980

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. doi:10.1177/0013164402239314

APPENDIX

SELECTED DERIVATIONS

Derivations for equation 37.

$$\mathrm{T} = \sum_{i=1}^{k} \tau_i$$

$$\mathrm{T} = \tau_1 + \tau_2 + \ldots + \tau_k$$

$$\operatorname{var}(\mathrm{T}) = \operatorname{var}(\tau_1 + \tau_2 + \ldots + \tau_k) + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \operatorname{cov}(\tau_i, \tau_j)$$

$$\operatorname{var}(\mathrm{T}) = \sigma_{\tau_1}^2 + \sigma_{\tau_2}^2 + \ldots + \sigma_{\tau_k}^2 + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{\tau_i \tau_j}$$

$$\operatorname{var}(\mathrm{T}) = \sum_{i=1}^{k} \sigma_{\tau_i}^2 + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{\tau_i \tau_j} = \sigma_{\mathrm{T}}^2$$

$$\sum_{i=1}^{k} \sigma_{\tau_i}^2 \geq \frac{\sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}}{k-1}$$

$$\sum_{i=1}^{k} \sigma_{\tau_i}^2 + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}} \geq \frac{\sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}}{k-1} + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}$$

$$\sum_{i=1}^{k} \sigma_{\tau_i}^2 + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{\tau_i \tau_j} \geq \frac{\sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}}{k-1} + \left(\frac{k-1}{k-1}\right) \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}; \quad \sigma_{x_{ij}} = \sigma_{\tau_i \tau_j}$$

$$\sigma_{\mathrm{T}}^2 \geq \frac{\sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}}{k-1} + \frac{k \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}} - \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}}{k-1}; \quad \sigma_{\mathrm{T}}^2 = \sum_{i=1}^{k} \sigma_{\tau_i}^2 + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{\tau_i \tau_j}$$

$$\sigma_{\mathrm{T}}^2 \geq \frac{k \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}}}{k-1}$$

$$\sigma_{\mathrm{T}}^2 \geq \frac{k}{k-1} \left(\sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{x_{ij}} \right)$$

Derivation for numerator of equation 48.

$$\mathrm{T} = \sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau \right)$$

$$\operatorname{var}(\mathrm{T}) = \operatorname{var}\left(\sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau \right) \right)$$

$$\operatorname{var}(\mathrm{T}) = \operatorname{var}\left(\sum_{j=1}^{k} \alpha_j + \sum_{j=1}^{k} \beta_j \tau \right)$$

$$\operatorname{var}(\mathrm{T}) = \operatorname{var}\left(\sum_{j=1}^{k} \alpha_j \right) + \operatorname{var}\left(\sum_{j=1}^{k} \beta_j \tau \right)$$

$$\operatorname{var}(\mathrm{T}) = \left(\sum_{j=1}^{k} \beta_j \right)^2 \sigma_\tau^2; \quad \operatorname{var}\left(\sum_{j=1}^{k} \alpha_j \right) = 0 \text{ as } a_j \text{ are constants}$$

Derivation for denominator of equation 48.

$$X = \sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau + \varepsilon_j \right)$$

$$\operatorname{var}(X) = \operatorname{var}\left(\sum_{j=1}^{k} \left(\alpha_j + \beta_j \tau + \varepsilon_j \right) \right)$$

$$\operatorname{var}(X) = \operatorname{var}\left(\sum_{j=1}^{k} \alpha_j + \sum_{j=1}^{k} \beta_j \tau + \sum_{j=1}^{k} \varepsilon_j \right)$$

$$\operatorname{var}(X) = \operatorname{var}\left(\sum_{j=1}^{k} \alpha_j \right) + \operatorname{var}\left(\sum_{j=1}^{k} \beta_j \tau \right) + \operatorname{var}\left(\sum_{j=1}^{k} \varepsilon_j \right)$$

$$\operatorname{var}(X) = \left(\sum_{j=1}^{k} \beta_j \right)^2 \sigma_\tau^2 + \sum_{j=1}^{k} \sigma_{\varepsilon_j}^2; \quad \operatorname{var}\left(\sum_{j=1}^{k} \alpha_j \right) = 0 \text{ as } a_j \text{ are constants}$$

APPENDIX

SOURCE CODE

This appendix includes source code for estimating credible intervals for the Bayesian coefficient alpha method presented in the manuscript. All code was written in R.

```
library(invgamma)
library(HDInterval)

bay_alp_jvd_g_v2= function(x, in_dat=0, alp_0=0, k_0=0, var_0=0, 
v_0=0, n_itr=1000, conf=.95){
#------------------------------------------#
# Author: J. V. DelosReyes                 #
# Written:  11/30/2022                     #
# Modified: 10/25/2023                     #
#------------------------------------------#
# R function to estimate Bayesian          #
# coefficient alpha (alp).                 #
#------------------------------------------#
# INPUT ARGUMENTS                          #
#------------------------------------------#
# x= input data                            #
# in_dat= type of data for x               #
#         0= x is a data matrix            #
#         1= x is a covariance matrix      #
# alp_0= prior alp                         #
# k_0=  prior sample size for alp          #
# var_0= prior alp variance                #
# nu_0= prior sample size for alp variance #
# n_itr = number of posterior draws        #
# conf= confidence level                   #
#------------------------------------------#
# OUTPUT                                   #
#                                          #
# return= Estimated alp, Bayesian alpha,   #
#         and CrIs                         #
#------------------------------------------#
```

```
cronbach_alpha<- function(x,in_dat=0) {
#-----------------------------------------#
# INPUT ARGUMENTS                         #
#                                         #
# x= input data                           #
# in_dat= type of data for x              #
#         0= x is a data matrix           #
#         1= x is a covariance matrix     #
#-----------------------------------------#
# OUTPUT                                  #
#                                         #
# return= Cronbach Alpha                  #
#-----------------------------------------#

  if (in_dat==0) {
    covx= cov(x)
  }

  if (in_dat==1) {
    covx= x
  }

  p= ncol(covx)
  x_num= sum(diag(covx))
  x_dem= sum(covx)
  return((p/(p-1))*(1-x_num/x_dem))

}

##-- Start bay_alp_jvd --##

# from data
n= nrow(x)
p= ncol(x)
cov_est= cov(x)
sum_cov_est= sum(cov_est)
cov_est_2= cov_est %*% cov_est
sum_cov_est_2= sum(cov_est_2)

alp_est= cronbach_alpha(x=x,in_dat=in_dat)
alp_var_est= ((p/(p-1))^2) * (2/(sum_cov_est^3)) *
  (sum_cov_est * (sum(diag(cov_est_2))+sum(diag(cov_est))^2) -
2*(sum(diag(cov_est)))*sum(cov_est_2))
```

```
# priors
# alp_0= 0
# k_0= 0
# var_0= 0
# v_0= 0

# posterior setups
k_n= n + k_0
v_n= n + v_0

var_n= chol((n-1)*alp_var_est + (v_0-1)*var_0 +
((n*v_0)/(n+v_0))*t(alp_est-alp_0)%*%(alp_est-alp_0))
alp_n= ((n*alp_est) + (k_0*alp_0)) / k_n

cnt_itr= 0
bd_jvd= 0
balp_post= matrix(data=0,nrow = n_itr,ncol=2)

while(cnt_itr<=n_itr){
  var_n_pull= rinvgamma(1, v_n, (v_n*var_n))
  balp_post_fit= rnorm(1, alp_n, (var_n_pull)/k_n)

  if(balp_post_fit<1){
    balp_post[cnt_itr,1]=cnt_itr
    balp_post[cnt_itr,2]=balp_post_fit
    cnt_itr= cnt_itr+1
  } else {
    bd_jvd=bd_jvd+1
  }
}

# output
balp_cri= t(quantile(balp_post[,2], c((1-conf)/2,
1-((1-conf)/2))))
names_cri= c("new_lcri","new_ucri")
colnames(balp_cri)=names_cri
balp_new_mn= mean(balp_post[,2])
balp_new_md= median(balp_post[,2])
balp_new_var= var(balp_post[,2])

# additional credible intervals
balp_hpdi= hdi(balp_post[,2], credMass = conf)
balp_hpdi= c(balp_hpdi[1],balp_hpdi[2])
hpdi_names= c("balp_l_hpdi","balp_u_hpdi")
balp_hpdi= t(as.matrix(balp_hpdi))
colnames(balp_hpdi)= hpdi_names
```

```
z_crit= -qnorm((1-conf)/2)
balp_new_se= sqrt(balp_new_var)
balp_z_lci= balp_new_mn - (z_crit * balp_new_se)
balp_z_uci= balp_new_mn + (z_crit * balp_new_se)
balp_z_ci= cbind(balp_z_lci,balp_z_uci)

balp_out=cbind(alp_est,balp_new_mn,balp_new_md,balp_new_var,
balp_cri,balp_z_ci,balp_hpdi,n_itr,n,bd_jvd,conf)
return(list(balp_cri=balp_out,balp_sim=balp_post))
}
```

VITA

John Mart V. DelosReyes

Department of Psychology
Mills Godwin Building, Room 250
Old Dominion University
Norfolk, VA, 23529

EDUCATION

Ph.D., Psychology (expected December 2023), Old Dominion University, Norfolk, VA
M.S., Psychology (December 2019), Old Dominion University, Norfolk, VA
B.S., Psychology (May 2015), Old Dominion University, Norfolk, VA

PUBLICATIONS

DelosReyes, J. M. V. & Padilla, M. A. (2023). Correlation coefficient confidence interval estimation via the bootstrap: The positive impact of a symmetric distribution. *Journal of Experimental Education.* doi: 10.1080/00220973.2023.2196659

DelosReyes, J. M. V. & Padilla, M. A. (*accepted*). Estimation of the correlation confidence intervals via the bootstrap: Non-normal distributions. *Journal of Modern Applied Statistical Methods*.

Zaharieva, J. N., DelosReyes, J. M. V., Cullen, K. K., & Padilla, M. A. (2014). *Do published student evaluations of teaching push faculty towards grade inflations and decreased quality of instruction?* Association for Psychological Science Student Research Grant Competition.

PRESENTATIONS

DelosReyes, J. M. V. & Padilla, M. A. (2023). *Investigating a new derivation for estimating Bayesian coefficient alpha*. 2023 Association for Psychological Annual Convention, Washington, D.C.

DelosReyes, J. M. V. & Padilla, M. A. (2022). *The positive impact of a symmetric distribution in correlation confidence interval estimation*. 2022 Association for Psychological Annual Convention, Chicago, IL.

DelosReyes, J. M. V. (2020). *Stats review*. Old Dominion University, Norfolk, VA.

DelosReyes, J. M. V. & Padilla, M. A. (2019). *Correlation confidence intervals robust to non-normality*. 2019 Association for Psychological Annual Convention, Washington, D.C.

DelosReyes, J. M. V. & Padilla, M. A. (2019). *Estimation of the correlation via the bootstrap: Non-normal distributions*. William and Mary Graduate Research Symposium, Williamsburg, VA.

DelosReyes, J. M. V. (2015). *How to read a scientific article: A primer into basic research literacy.* Psi Chi International Honor Society.

DelosReyes, J. M. V., Manning, S., Sabo, S., Deshpande, A., & Padilla, M. A. (2015). *Developing a measure of psychological aggression: First steps*. Virginia's Collegiate Honors Council Conference, Richmond, VA.