When what is wrong seems right: A Monte Carlo simulation investigating the robustness of coefficient omega to model misspecification

Stephanie Bell

A thesis submitted to the Faculty of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Arts

Graduate Program in Psychology

York University

Toronto, Ontario

August 2021

Abstract

Coefficient omega is a model-based reliability estimate that is unrestricted by assumptions of a unidimensional essentially tau equivalent model. Rather, omega can be adapted to suit the underlying factor structure of a given population. A Monte Carlo simulation was used to investigate the performance of unidimensional omega and omega-hierarchical under circumstances of model misspecification for high and low reliability measures and different scale lengths. In general, bias increased with the amount of unmodeled complexity (i.e. unspecified multidimensionality or error correlations). When models were misspecified, observed bias was higher when true population reliability was lower, and increased with scale length. Less variable estimates were observed when true reliability and sample size were higher.

Table of Contents

List of Tables

List of Figures

Introduction

Reliability estimates provide information about how well an observed score represents the construct being measured. Although coefficient alpha (Cronbach, 1951; Guttman, 1945) has long reigned as the most commonly reported reliability statistic for composite score reliability, the assumptions on which it is based are easily violated, resulting in biased estimates of true reliability (e.g. Dunn et al., 2014; Graham, 2006; Green & Yang, 2009; Raykov, 1997). In particular, multidimensionality can result in strong bias and render coefficient alpha estimates uninterpretable (Stanley & Edwards, 2016). Instead, a more accurate and interpretable estimate can be obtained by using the parameters of a confirmatory factor analysis (CFA) to calculate model-based reliability in the form of coefficient omega (McDonald, 1999). However, true population models cannot be known, and researchers must rely on a variety of statistical methods, previous evidence, and theory support their theorized structures. Simulations have suggested that goodness of fit statistics may be biased in favour of certain models, such as the bifactor model (e.g. Bonifay & Cai, 2017; Reise et al., 2016). As a result, the CFA could be misspecified, and the estimate of omega based on incorrect values. As yet, the degree to which different coefficient omega statistics differ from true reliability when the model is incorrectly specified is not known. Using a Monte Carlo simulation, this thesis will investigate the performance of coefficient omega for unidimensional models and omega-hierarchical in cases of misspecified models. Additionally, both estimates will be calculated using the observed variance as well as the model-implied variance to test the assertion that these results are equal (Kelley & Pornprasertmanit, 2016) when the model is incorrect.

Classical Test Theory and Reliability

Following from Classical Test Theory (CTT; Lord & Novick, 1968), for any test

component (i.e., item) *j*, the observed score (x_j) is presumed to be equal to the sum of an individual *i*'s true score for the item (τ_j) and error (ε_j). For any test that can be represented by a total composite score *X*, it follows that

$$X_i = \sum_{j=1}^{J} \tau_{ij} + \sum_{j=1}^{J} \varepsilon_{ij}$$

Here, the true score represents an individual's score for systematic influences on the test (such as the underlying construct). Conversely, the error term captures noise left over, which may positively or negatively impact the individual's observed score relative to their true score. Theoretically, across infinite parallel test items, this error should cancel to 0, leaving the observed and true total scores to be equal. However, for any individual test, there will always be some degree of error. The amount of error relative to true score, which is quantified by reliability, must be understood to evaluate the usefulness of a given measure. Population reliability of a composite score can be defined as

$$\rho(X) = \frac{\sigma_\tau^2}{\sigma_X^2}$$

where σ_τ^2 represents the variance due to total true score and σ_X^2 represents the overall composite score variance. Reliability therefore represents the proportion of total variance which is attributable to the true score. However, the challenge faced in practice is that neither true score nor error can be directly observed. As a result, reliability can only be approximated, and there are many different methods to do so.

Coefficient Alpha

Coefficient alpha (α), also known as Cronbach's alpha (Cronbach, 1951; Guttman, 1945), is the most common reliability estimate and, in many cases, the only estimate reported (Flake et al., 2017). The ubiquity of coefficient alpha may be part of its appeal, substituting ease for

precision (Black et al., 2015). Unlike alternative measures of reliability discussed later, coefficient alpha can be calculated from a single test administration and requires only the computation of a covariance matrix or item variances and an unweighted total score. Reliability analysis from statistical software such as SPSS will report coefficient alpha as its main or only output, and few applied researchers are aware of its alternatives (Black et al., 2015; Yang & Green, 2011).

Coefficient alpha represents the mean split-half correlation for every possible split of items for a given scale adjusted for scale length (Cronbach, 1951). With a single administration of a scale, assuming item errors are uncorrelated, a lower-bound estimate of reliability can be calculated from

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} \sigma_{y_j}^2}{\sigma_X^2}\right)$$

where J is the number of items in the scale, $\sigma_{y_j}^2$ is the variance for item j, and σ_X^2 is the composite score variance. This formula may be expressed equivalently as

$$\alpha = \frac{J^2 \bar{\sigma}_{jj'}}{\sigma_X^2}$$

where $\bar{\sigma}_{jj'}$ is the average covariance among the items (Green & Yang, 2015; McDonald, 1999).

The formula for coefficient alpha is based on the assumption of *fungible units*, which treats any one item as interchangeable with any other item without changing scale properties. This assumption can be relaxed to take the form of the *essentially tau-equivalent* model, a factor analytic model wherein all items load equally onto a single factor representing a single underlying construct. Items are free to vary in intercepts, but the slopes remain the same. Slopes, in the case of a factor analytic equation, represents the factor loadings of items, which quantify how the underlying construct represented by a factor influence items on a scale. Thus, in the tau-

equivalent model, all items are directly influenced in the same way by a single common factor. The equation can be centered around 0 to negate the intercept. Thus, the equation for the total score of a one-factor tau-equivalent model can be expressed as

$$X_i = \sum_{j=1}^{J} \lambda f_{ij} + \sum_{j=1}^{J} e_{ij}$$

where *X* is the total observed score of individual *i* for a test comprised of *j* items, λ is the factor loading for all items onto the single factor, f_i is the factor score for individual *i*, and e_i is the error term associated with the items. The true score for any individual is thus the product of the factor loading and their own factor score. Factor score variance can be set to 1 in order to define the factor's metric, and coefficient alpha can be calculated using the parameters from the tau-equivalence model such that

$$\alpha = \frac{J^2\lambda^2}{\sigma_X^2}$$

where λ is the factor loading for all items onto a single factor (Green & Yang, 2015).

However, the tau-equivalence model demands important assumptions that must be met for coefficient alpha to be an unbiased estimate of true reliability: (1) the scale must be unidimensional such that all items are influenced by a single underlying construct, (2) factor loadings must be equal, such that all items are influenced in the same magnitude and direction, and (3) the errors are independent across items. Violations of these assumptions can result in a biased estimate of true reliability.

Essential tau equivalence is unlikely to be true under any real-world condition. Even when a scale is truly unidimensional, it is improbable that each item will have the same relationship with the factor. Factor loadings often vary widely, and this variance will result in underestimation of reliability by coefficient alpha (Graham, 2006; Green & Yang, 2009; Raykov,

1997). Simulations have shown this bias may be as strong as 11.1% under severe violations (Raykov, 1997). However, Savalei and Reise (2018) point out that the worst biases are likely to occur for short scales where one item has a high loading while the others are low, and that these scales are likely to already have low reliability. When other assumptions are met, coefficient alpha can serve simply as a lower-bound reliability estimate (Lord & Novick, 1968).

Violations of unidimensionality and uncorrelated errors can result in much more troubling consequences. Error covariances among items will produce bias in the direction of the covariance with a magnitude relative to its strength (Fleishman & Benson, 1987; Maxwell, 1968; Raykov, 2001). Yang and Green (2011) posit that residual correlations between items tend to be positive, producing an over-estimate of reliability. This effect was demonstrated in simulations by Gu et al. (2013), which indicated that the presence of correlated errors could result in an overestimate of population reliability by as much as .38 for a low reliability (ρ = .38) scale when 2/5 of items had residual correlations of .40. As true reliability increased, the bias of coefficient alpha decreased; however, even when the true reliability was .94, coefficient alpha overestimated reliability by .02, on average. Indeed, as Yang and Green (2011) predicted, as error correlations increased, so did the bias of coefficient alpha (Gu et al., 2013). Failing to address violations of the assumption of uncorrelated errors can therefore have serious repercussions. Model fit should always be considered prior to estimating reliability to assess residual covariance and possible multidimensionality.

Multidimensionality

Dimensionality refers to the number of systematic influences on a scale's score. A unidimensional scale, as discussed above, is one in which a single factor influences all items directly. However, defining a scale with a single factor is not necessarily equivalent to defining a

scale which measures a single construct. Several authors have argued that only scales with a very narrow scope will satisfy the assumption of unidimensionality and most scales show some degree of multidimensionality (Reise et al., 2007; Reise et al., 2000; Socan, 2000). A single construct that influences all items in a scale may still be present among multiple factors.

Multidimensionality typically occurs for one of two reasons. A test may be designed to measure multiple distinct factors related to an overarching construct, or methodological artefacts may result in nuisance factors (Yang & Green, 2011). In the case of the former, these factors may present as subscales or variance related to constructs of interest (e.g. inattention-related variance and hyperactivity-related variance for a measure of ADHD). In the case of the latter, factors can result from residual covariances produced by item wording effects, multiple items pertaining to the same stimulus, and so on.

Multidimensionality may manifest in a number of forms. Bifactor models have become increasingly popular, with applications in major areas of psychological research, including personality, intelligence, and psychopathology (Rodriguez et al., 2016). In a bifactor model, a single general factor directly influences all items in a scale, while specific factors capture residual covariances that remain over and above the general factor for smaller subsets of items. For model identification and interpretability of factors, the general and specific factors should be orthogonal to one another (Reise, 2012; Rijmen, 2009).

An example of a path diagram for a bifactor model can be seen in Figure 1. Each item in the bifactor model is related to the general factor and one specific factor with error variance unique to that item. Specific factors may capture residual variance related to meaningful constructs or simply account for the effects of methodological artefacts. The general factor is intended to represent a single overall construct influencing all items in a scale. In this way, the

model allows for a multidimensional measure of a single construct, improving fit while potentially retaining a meaningful composite score. We can assess the appropriateness of composite score use by calculating the reliability of the composite score as a measure of the construct represented by the general factor, using a form of model-based reliability, omega-hierarchical, which will be discussed later.

Alternatively, multidimensionality may present as a higher-order or multiple factor model, both of which are nested within the equation for a bifactor model (Reise, 2012). Higher-order models are similar to bifactor models in that all items are influenced by a single general factor, as well as one of several lower-order factors. However, in a bifactor model, the general and specific factors are orthogonal, while in a higher-order model, the general factor (or higher-order factor) indirectly influences the items through lower-order factors. Conversely, a general factor may not be present to create a multiple first-order factor model. In this case, two or more constructs represented by distinct, but correlated, factors directly influence subsets of items. Typically, each item is influenced by a single construct represented by one factor, but there may be cases of cross-loadings, wherein an item may be influenced by more than one construct.

Coefficient alpha is not appropriate for assessing the composite score of a test with an underlying multidimensional model, as multidimensionality will obscure interpretation of a reliability estimate that fails to distinguish the source of the variance. High values of coefficient alpha are often considered by applied researchers to indicate unidimensionality or appropriate use of composite score. In reality, both high and low values of alpha are associated with unidimensional or multidimensional tests (Schmidt, 1996; Sijtsma, 2009), and therefore coefficient alpha is not reflective of factor structure. Rather, the true score in the numerator of the equation for coefficient alpha is comprised of the average covariance among scale items and

thereby reflects all systematic influences on an individual's test score. That is, coefficient alpha estimates the proportion of variance due to all systematic influences, and does not necessarily reflect the proportion of variance related to the construct of interest. If one is interested in the proportion of variance due to a particular construct, this quantity would be better estimated with a model-based reliability estimate.

Coefficient Omega

Coefficient omega (ω; McDonald, 1999) is an alternative, model-based reliability estimate, which is based on the congeneric model. The congeneric model can be expressed as

$$x_{ij} = \lambda_j f_i + e_{ij}$$

where x_{ij} is the observed score for individual *i* on item *j*, λ_j is the factor loading for each item *j*, f_i is the factor score for any individual *i*, and e_{ij} is the error term for the item. Within the congeneric factor model, factor loadings are allowed to vary across items. In this way, the congeneric model is represented by a regression equation in which each item is regressed onto a single factor with a slope equal to its factor loading. True score here is the product of the factor loading and the individual's factor score, so that the total true score for any individual is then the sum of the estimated true scores for each item.

Coefficient omega, based on the parameters of this model, represents the proportion of observed composite score variance attributable to the single factor. Generally, CFA is preferred to exploratory factor analysis when estimating coefficient omega, as a CFA implies a stronger theory for the underlying model (Flora, 2020). There is some evidence from simulations that reliability estimates are more accurate when a CFA approach is used in cases for which the underlying model includes a single factor common to all items (Murray et al., 2019). To calculate coefficient omega for a unidimensional test, we can set the factor score variance equal

to 1 to set the metric so that

$$\omega = \frac{(\sum_{j=1}^{J} \lambda_j)^2}{(\sum_{j=1}^{J} \lambda_j)^2 + \sum_{j=1}^{J} var(e_{ij})}$$

where$\sum_{j=1}^{J} var(e_{ij})$ represents the total error variance across all items. The denominator of coefficient omega can be represented as the model-implied total variance as seen above or as the observed total variance (σ_X^2) without any meaningful difference, though Kelley and Pornprasertmanit (2016) posit observed total variance may be more robust in the case that a model is misspecified. If the denominator is to be calculated as the model-implied variance and there are residual covariances for an otherwise unidimensional model, the denominator should be expanded to include an additional term of two times the sum of the residual covariances. Coefficient omega can also be modified to accommodate categorical items using CFA parameters based on polychoric correlations (Green & Yang, 2009); however, this modification will not be explored here.

Coefficient omega may also be modified to provide interpretable reliability estimates for tests with multidimensional measurement models. For a scale well-represented by a bifactor model, omega-hierarchical (ω_H) is a suitable composite score reliability estimate (Zinbarg et al., 2005). The formula for omega-hierarchical is nearly the same as above but for a few modifications. The numerator includes the factor loadings for only the general factor (λ_{gj}) and, if the denominator is represented as the model implied total variance, it must expand to include terms for the squared sum of factor loadings for all specific factors (λ_{snj}) as well as the general factor and error. The formula can be written as

$$\omega_H = \frac{(\sum_{j=1}^{J} \lambda_{gj})^2}{(\sum_{j=1}^{J} \lambda_{gj})^2 + (\sum_{j=1}^{J} \lambda_{s_k j})^2 + \sum_{j=1}^{J} var(e_{ij})}$$

where $\lambda_{s_k j}$ is the factor loading of item j onto specific factor k. Omega-hierarchical thus represents the proportion of composite score variance that can be reliably attributed to the general factor. If the reliability due to the general factor is high, then much of the variance in test scores is caused by a single underlying construct, and the scale can be treated as "essentially unidimensional" (Rodriguez et al., 2016). Thus, the use of a composite score for the scale would be appropriate.

The Problem with Bifactor Models

Despite their advantages, methodologists caution against overusing bifactor models (e.g. Bonifay et al., 2017; Markon, 2019; Maydeu-Olivares & Coffman, 2006; Reise et al., 2016). Although fit statistics may favour the bifactor structure over other models, this may be a symptom of overfitting, rather than a representation of the true underlying structure. Reise et al. (2016) found that the bifactor model was able to produce an adequate fit for the Rosenberg Self-Esteem Scale (Rosenberg, 1965), not necessarily because of a "reverse-wording" effect, but because of a greater ability to capture invalid response patterns. Bonifay and Cai (2017) demonstrated that even when data were generated to follow random patterns, the bifactor model showed a good fit for a high percentage of samples. Simulations have shown that a bifactor model can produce as good or better fit statistics than the correct model when fit to data from unidimensional, two-factor, and higher-order populations (e.g. Maydeu-Olivares & Coffman, 2006; Morgan et al., 2015; Murray & Johnson, 2013). The bifactor model can appear to fit even when it is inappropriate.

In real-world situations, we can only approximate the true model underlying any measure, using fit statistics to indicate how well a hypothesized model fits sample data. A bifactor model which produces good fit to data may then be selected even though the true model

is not bifactor. In that case, omega estimates will be based on a bifactor model, and there is a risk that omega may be biased as a result. Currently, little is known regarding the accuracy of unidimensional omega and omega-hierarchical given incorrect model selection. Using a series of Monte Carlo simulations, this study investigated the following research questions:

1. How well does omega for a unidimensional model estimate composite score reliability when the true model is not unidimensional?
2. How well does omega-hierarchical perform as an estimate of composite score reliability when the true model is not a bifactor model?
3. To what extent is the degree of bias related to goodness of fit statistics used to assess model fit?

I hypothesized that coefficient omega for unidimensional models and omega-hierarchical will both experience some degree of bias when calculated for models that are incongruent with their design. I expected that the extent of the bias would depend on the degree to which CFA parameters differ from their true model, or the degree to which factor loadings have been affected by model misspecification. For models with a general factor, I hypothesized that these coefficients will have greater accuracy when the general factor of a given model is strong. That is, when factor loadings for the general factor tend to be high, such that it accounts for a large proportion of variance, bias of both unidimensional omega and omega-hierarchical would be smaller than for scales with a weak general factor. Models with poor fit are unlikely to be selected, so the relationship between model fit and bias was assessed for a clearer picture of bias among models that would be more likely to be chosen based on fit. Finally, I will be investigating whether bias is affected by the use of the observed or model-implied total variance denominator.

Method

To investigate the bias of unidimensional omega and omega-hierarchical, a series of Monte Carlo simulations were run using R (R Core Team, 2020). Data were drawn from multivariate normal distributions with covariance structures consistent with given population CFA models using the mvrnorm function of the MASS package (Venables & Ripley, 2002). Sample models were estimated using the maximum likelihood estimator in the lavaan package (Rosseel, 2012) and reliability was estimated using semTools (Jorgensen et al., 2020). True reliability was calculated for each population model and compared with sample estimates of reliability for 1,000 random samples for each cell of the study design. In total, there were four population model factor structures and both a high and low reliability model were generated for both long and short scales. For each condition, reliability was assessed for three sample sizes, creating a total of $(4 \times 2 \times 2 \times 3) = 48$ unique cells.

Study Conditions

For each of the models described below, there were two conditions of reliability, determined as a function of factor loadings within the population model. The high reliability condition set reliability to approximately .85, while the calculated reliability of the low reliability condition was approximately .60. Scale lengths were either short (8 items) or long (16 items). To ensure enough indicators per factor for model identification, scale lengths were necessarily longer for the higher order model, such that the short test was 12 items (3 indicators per lower-order factor), and the long test was 20 items (5 indicators per lower-order factor).

Sample sizes were large (N = 1,000), medium (N = 250), and small (N = 100). Often, a sample size of 100 would not be considered suitable for factor analysis; however, in applied cases when scale structure and psychometric properties are not the focus, researchers may still

wish to estimate the reliability of their selected measures. In these cases, samples may be small relative to psychometric studies. Sample sizes of 1,000 are more ideal for factor analytic purposes, but are often unrealistic in practice, especially in applied studies.

Sample data were drawn from four population models: a simple one-factor model with no correlated errors, a one-factor correlated errors model, a bifactor model, and a higher-order model. All models were specified such that factor variances equaled 1.0; consequently, the population-level model-implied covariance structures were in the correlation metric. Although unlikely to occur in real-world scenarios, the simple one-factor model was included to investigate the performance of unidimensional omega under ideal conditions. Samples from these population models were fit only to the correct model across replications. For all remaining population models, a simple one-factor model, a correlated errors model, and a bifactor model were fit to sample data. Therefore, data from the correlated one-factor and bifactor population models were fit to a correctly specified model as well as two incorrectly specified models, while data from the higher-order populations were fit only to incorrectly specified sample models.

Figures 2 through 5 show the structure of the correct population models in the shorter scale length condition. Table 1 shows the complete list of population factor loadings for each population model. Population models were not designed to be tau equivalent; instead, factor loadings differed across items. Factor loadings for the correlated one-factor model ranged from .493 to .837 in the high reliability conditions and from .493 to .624 in the low reliability conditions. Because allowing all item errors to correlate would have produced under-identified models, the one-factor model with correlated errors was specified such that half of the items correlated with one another. The error correlations were small to moderate in the high reliability conditions (approximately .09 to .31) and moderate to high in the low reliability conditions

(approximately .19 to .52). In a typical bifactor model, every item loads onto both a general factor and a specific factor. However, preliminary simulations showed that bifactor models estimated using data from the correlated errors population could not converge consistently with two specific factors. Therefore, only one specific factor representing these error correlations was included in addition to the general factor.

The bifactor population models were specified to include two specific factors pertaining to equal halves of the items and a single general factor underlying all items. In the high reliability conditions, general factor loadings ranged from .358 to .913, while specific factor loadings were smaller and ranged between .213 and .448. For low reliability conditions, general factor loadings ranged between .314 and .711, while specific factor loadings ranged between .213 and .663. The correlated errors model fit to sample data allowed all items within each half to correlate with one another, but not with items from the other half of the scales.

Finally, the higher-order population model included a single higher-order factor and four lower-order factors. Four was selected as the number of specific factors because this is the smallest number of factors from which a higher-order model can be empirically distinguished from a correlated-factor model (Rindskopf & Rose, 1988). For the higher-order factor, loadings ranged between .72 and .91 in the high reliability condition and .51 to .69 in the low reliability condition. Lower-order factor loadings ranged from .68 to .91 for the high reliability condition and .44 to .79 in the low reliability condition.

Evaluation of Results

For each replication, unidimensional omega and omega-hierarchical were calculated and compared with true population reliability. Unidimensional omega was calculated for both the correlated error and uncorrelated error models and were analyzed separately. Omega estimates

using both model-implied and observed total variance were included in the analysis. Bias for each estimate within each condition was calculated as the difference between the omega estimate and true reliability ($\omega - \rho$). Precision was assessed as the variability within each cell represented by the standard deviation. Finally, the relationships between mean bias and goodness of fit statistics of CFI, TLI, and RMSEA were plotted and correlated to determine whether higher bias was only a concern in cases of poor model fit, or whether it could frequently occur in other conditions.

Results

Convergence and Proper Solutions

Convergence and proper solution rates for each cell are shown in Table 2. Across all conditions and replications, 98.07% of the sample models converged. All except one of the nonconverged models was from a bifactor sample model. Of the estimated models that converged, approximately 95.11% produced proper solutions. Bifactor models also produced more improper solutions, except when the true model was bifactor and reliability was low. In this case, the correlated errors model produced far fewer proper solutions. In general, convergence and proper solution rates increased as sample size and true reliability improved. Solutions that did not converge and produce proper solutions were excluded from analysis.

Simple One-Factor Population Model

Table 3 shows the mean and standard deviation of bias of unidimensional omega correctly specified to a one-factor model with no correlated errors. Unidimensional omega showed an average bias of approximately 0 for all conditions, except when true reliability was low (ρ = .60) and sample size was small (n = 100). These conditions produced a small bias, overestimating reliability by approximately .01. Mean bias did not exceed .01 in either direction.

Precision, as expressed by the standard deviation of absolute bias, was most strongly related to sample size and true reliability, such that there was less variability among estimates as sample size increased and when true reliability was higher. The distributions of bias for each condition within each sample size are shown in Figure 6. Sample sizes of 100 yielded a standard deviation of .02 when true reliability was high; however, variability increased to .07 in the low reliability condition. As sample size increased to 250, precision improved to .01 for samples from a high reliability population and .04 for samples from the low reliability population.

Overall, given a reasonable sample size, unidimensional omega produced a good estimate of reliability in the theoretical ideal circumstance of a one-factor model with independent errors.

One-Factor Population Model with Correlated Errors

The mean bias and variability for omega estimates for a one-factor population with correlated errors are shown in Table 4. Similar to the correctly specified unidimensional omega in the previous section, a correct model yielded unbiased estimates on average, regardless of condition. The variability of absolute bias also followed a similar pattern to above, such that the distribution of estimates became narrower as sample size and true reliability increased. The worst precision was seen for small sample sizes with low reliability, for which standard deviations were .09 and .07 for the 8- and 16-item measures, respectively. Variability was still rather high when sample size was 250, but improved to .06 and .05, respectively. High reliability conditions yielded more precise estimates, such that even small sample sizes had a standard deviation of bias around .03. Boxplots for each condition within each sample size are in Figures 7 through 9.

The misspecified bifactor model yielded similar results to the correct model, such that high reliability conditions yielded the same relatively unbiased results. However, in low true reliability conditions, omega-hierarchical produced an average bias between -.01 and .02. Precision was also slightly worse in the 16-item low reliability for omega-hierarchical, but only increased by a difference of .01 relative to unidimensional omega. Specifying a bifactor sample model therefore produced worse estimates than the correct model, but the overall difference was small. In general, omega-hierarchical produced reasonable estimates when misspecified to data from a one-factor population with correlated errors.

The misspecified simple one-factor model produced highly biased estimates, on average. Absolute bias was the smallest for scales with fewer items and high true reliability, but even in

the best of circumstances, still ranged between .06 and .08. In low reliability 16-item conditions, bias ranged between .22 and .25. The variabilities of these estimates were small relative to the other sample models, but given that the standard deviations were centered around such a high bias, unidimensional omega showed a poor performance when the model failed to include correlated errors.

Bifactor Population Model

Table 5 shows the mean absolute bias and variability of omega estimates when the population model was bifactor. The correctly specified bifactor sample model produced omega-hierarchical estimates which were relatively unbiased, although a small overestimation of .01 appeared in some low reliability conditions. Precision showed a similar pattern as with the previous population models, such that there was low variability in estimates from high reliability populations, but variability in estimates from low reliability was about .08 when sample size was 100 and dropped to .05 as sample size increased to 250. Figures 10 through 12 show the distributions for each condition within each sample size. Omega hierarchical was therefore a good estimate of reliability for the bifactor model, assuming sufficient sample size for low estimate variability.

For high reliability conditions, misspecification of a unidimensional model with correlated errors to the bifactor population produced relatively unbiased estimates of reliability. However, bias in the low reliability conditions increased to .01 for the 8-item condition and up to .07 for the 16-item condition when sample size was low. As sample size increased, bias dropped to .05. Estimates showed less variability for unidimensional omega relative to omega-hierarchical; however, the risk of bias was higher overall, indicating that it is not a suitable replacement for omega-hierarchical.

As with the one-factor with correlated errors population, the simple unidimensional sample model produced estimates with very high bias, ranging from .08 in the high reliability condition to .28 in the low reliability condition. Mean bias increased as true reliability decreased and as number of items increased. There was low variability in these estimates, but the simple model was too strongly biased in all conditions to be an acceptable alternative to omega-hierarchical.

Higher-Order Population Model

For the higher-order population, estimates based on the simple one-factor sample model displayed a similar pattern of high bias as with the other complex models. Absolute bias ranged from .08 to .23 across conditions, worsened by both scale length and low population reliability. However, specifying correlated errors improved the performance of unidimensional omega such that mean absolute bias ranged from -.04 to 0. Unidimensional omega was more likely to underestimate reliability when the sample size was small and true population reliability was low. As the sample size improved to 250, mean bias improved to near 0. Variability of estimate bias was once again about .09 when sample size was only 100, but reduced to approximately .05 given an increase in size to 250. Omega-hierarchical produced almost the same pattern as unidimensional omega with specified correlated errors. Bias was small, and for sample sizes 250 or higher, precision was acceptable even in populations with low reliability. Figures 13 through 15 show boxplots of omega estimates for each sample model for a given sample size. With a reasonable sample size, omega-hierarchical and unidimensional omega with specified correlated errors both produced reasonable estimates of true reliability in a higher-order population.

Denominators of Coefficient Omega

For all conditions, omega was calculated using both the model-implied total variance

(ω_Σ) and observed total variance (ω_S). In general, only small differences in omega estimates were observed between the two calculations. Differences in mean bias did not exceed .03, while differences in precision (estimate standard deviations) did not exceed .02. Typically, ω_Σ produced slightly higher and less variable reliability estimates. Accordingly, ω_S was more likely to underestimate omega, but this result was only apparent when fitting bifactor or correlated errors models to a higher-order population model. Differences between the two calculations were most common when the sample model was misspecified, and strongest for the worst misspecifications (i.e. estimating unidimensional omega based on a simple one-factor sample model when the true model was more complex). Differences were also related to low population reliability and were magnified as the number of items increased. In general, there was not a meaningful difference between the estimate calculations.

Relationship with Model Fit

To investigate how model fit related to the bias of coefficient omega, scatterplots were graphed and correlations were calculated between mean bias and three fit statistics (RMSEA, CFI, and TLI). Given that mean bias could be positive or negative, absolute values of bias were used for this analysis. Fit statistics were individually regressed onto absolute bias to assess characteristics of each relationship. Plots indicated that the overall relationships were linear, but residuals were heavily positively skewed, so correlations were calculated using Spearman's rank-order correlations. Additionally, two replications were identified as potential outliers with high Cook's distance for TLI. These values were both from the low reliability simple one-factor population with 16 items and a sample size of 100. TLI values for these replications were -10.56 and 13.14, respectively, and had corresponding CFI values of 1 and RMSEA of 0. Removal of these cases did not meaningfully change the observed relationships between fit statistics and

absolute bias. Additionally, removal of all cases where TLI exceeded 1.10 did not meaningfully change these results and were therefore retained.

Figures 16 through 18 show the average relationships between fit statistics and bias for each sample within each model. Although the overall trend did not appear to violate linearity, plotting mean bias against CFI or TLI for different sample models within each population revealed more curvature in an upside-down parabola or wave-like function which appeared to be a product of combining high- and low-reliability conditions. Given poor fit of models where most of the curvature occurs, it is unlikely these models would be selected for use. Curvature was therefore not investigated further. Where fit statistics would typically indicate good model fit (approaching 1 for CFI and TLI; approaching 0 for RMSEA), variability of bias was high, but the average absolute bias changed only by about .01 to .02.

The relationships between absolute bias and all three fit statistics for the whole dataset were strong (r_{RMSEA} = .58, r_{CFI} = -.64, r_{TLI} = -.61), suggesting that poorer model fit was associated with greater bias. However, for each sample model, worse fit did not appear to consistently increase bias, supporting that these correlations may be a function of merging the data from all three population models. In general, fit statistics and bias were both poor for unidimensional omega, while unidimensional omega with specified correlated errors and the bifactor model both showed low bias and good fit. Therefore, merging the two supports a stronger relationship than may truly be present within each model. Table 7 shows the correlations between model fit and bias for each fit statistic within population-level and sample-level data. Generally, CFI showed the strongest relationship with bias. Associations were strongest for unidimensional omega with no specification of correlated errors, and weakest for the omega-hierarchical, regardless of population.

Discussion

Methodologists have suggested transitioning from coefficient alpha to model-based reliability estimates such as coefficient omega which consider the underlying model and therefore do not rest on assumptions of unidimensionality and essential tau equivalence (e.g. Flora, 2020; Raykov & Marcoulides, 2011). In the current study, Monte Carlo simulation techniques were used to investigate the bias of unidimensional omega and omega-hierarchical when the sample model is misspecified. Omega estimates based on sample models correctly specified to their respective population model were unbiased on average (mean absolute bias between -.01 and .01), regardless of other sample or population characteristics. Given a reasonable sample size ($N \geq 250$), omega also showed a reasonable estimate variability, suggesting that it performs well as an estimate of reliability when the model is correctly specified. Conversely, misspecified models produced mean biases ranging from -.02 to .28. Consistent with hypotheses, the degree to which estimates of omega were biased by misspecification of the sample models was primary related to the amount of unmodeled complexity in the sample model. That is, unidimensional omega based on a simple one-factor sample model consistently performed poorly for complex models, but as correlated errors and multidimensionality were specified, performance improved. Selection of observed versus model-implied total variance for the calculation of omega estimates did not have much impact on results or improve bias in the case of misspecification, contrary to the claim by Kelley and Pornprasertmanit (2016) that use of observed total variance is more robust to misspecification.

Unidimensional omega based on a one-factor sample model with no correlated errors was a suitable reliability estimate only when the sample model was correctly specified. Estimates for data from more complex populations were highly biased, averaging .07 or higher, even under the

best of circumstances. Specification of correlated errors improved unidimensional omega's performance, such that it was somewhat more robust against bias when the model was incorrectly selected. Estimates were relatively unbiased when true population reliability was high, indicating a greater proportion of variance was related to a single construct, or when tests had fewer items. Low reliability conditions for longer tests showed that unidimensional omega was biased, on average, by .04 if the data was from a higher-order population, and up to .07 if the data was from a bifactor population. In these situations, the specification of correlated errors for all items within specific factors did not appear to be enough to capture multidimensionality. For shorter measures, it appears that unidimensional omega may be estimated with minimal risk of bias if error correlations are specified, but this is not true for longer measures. To minimize bias risk, unidimensional omega should only be used when there is sufficient evidence to suggest a truly unidimensional measure.

Compared with unidimensional omega with or without correlated errors, omega-hierarchical displayed the most stable performance between population model conditions. Previous simulations have indicated that the bifactor model may be overused and can produce good fit statistics despite misspecification due to overfitting (e.g. Maydeu-Olivares & Coffman, 2006; Morgan et al., 2015; Murray & Johnson, 2013). In cases where the bifactor model converged and produced proper results, omega-hierarchical showed little or no bias on average, even when the true model was not bifactor. As with other sample models, mean bias was highest in low reliability conditions, and this effect was exacerbated by longer scales. However, bias did not exceed .02 in either direction for any condition and, given a reasonable sample size, showed low variability. Thus, omega-hierarchical appeared largely robust to model misspecification. For any model with a single factor common to all items, incorrect selection of a bifactor sample

model will not badly bias omega estimates.

Results from these simulations are largely in alignment with previous studies, indicating that omega estimates are generally accurate when the sample model is consistent with the true model, and that reliability estimates are more likely to be biased in cases of misspecification as true reliability decreases (e.g. Gu et al., 2013; Yang & Green, 2010; Zinbarg et al., 2005). Scale length further amplified observed bias in misspecified models. A simulation by Yang and Green (2010) found that longer measures may have the effect of decreasing bias of omega for misspecified models by including more indicators; however, this finding was based on models which included zero to two correlated errors, or for a true bifactor model with only a single factor. Here, the degree of misspecification here was higher, such that the bifactor and higher-order populations included two specific and four lower-order factors, respectively, while the one-factor population with correlated errors included correlations for half of the sample items. More indicators therefore produced more complexity, thereby overriding the stabilizing effects of increasing indicators. The effect of scale length then appears to be related to the degree of misspecification.

These findings further support Yang and Green's (2010) observation that failure to specify correlated errors for a congeneric model will result in worse bias, and expand on the performance of model-based estimates for a bifactor sample fit to data from a population other than bifactor. In Yang and Green's (2010) simulation, misspecification of a congeneric model as bifactor was found to increase the risk of bias. However, their study did not fit the bifactor model to a population with correlated errors, so the degree of misspecification was higher. Such bias did not appear to the same extent in this study when the true model included correlated errors. Similarly, their study employed an SEM estimate more akin to omega-total, which includes a

variance term for any specific factors in its numerator, in addition to the general factor. It is therefore a measure of all systematic variance and is likely to produce higher estimates. Yang and Green (2010) hypothesized that the true bias associated with overspecified models was a symptom of overfitting and inflation of the true score estimate. Conversely, omega-hierarchical includes only the true score estimate related to the general factor, which may have removed some of the bias associated with misspecifying a congeneric model as bifactor.

This study did not directly compare the performance of omega estimates with coefficient alpha; however, a comparison of results from previous simulations with results presented here indicates that coefficient omega is less biased than coefficient alpha overall, but only when error covariances and multidimensionality are included in the CFA model. Gu and colleagues (2013) found that when residuals were correlated at .40 for 6 of 15 items, coefficient alpha overestimated reliability by .05 when the true reliability was .86, and by .18 when true reliability was .68. In the present study, the correlated errors model included error covariances between .20 and .50 for a 16-item test. When those error covariances were not modeled, unidimensional omega overestimated reliability by .09 when true reliability was .85, and by .22 when true reliability was .60. Yet, when the error covariances were correctly modeled, omega showed little to no bias on average. Given the differences in model specifications, results are not directly comparable; however, this pattern does suggest that coefficient omega is an improvement to coefficient alpha when the true model is complex, but only when that complexity is adequately modeled. This conclusion is more directly supported by Yang and Green's (2010) report that SEM estimates based on underspecified congeneric models (i.e. unidimensional omega based on a model without specified correlated errors) had a slightly higher relative bias than coefficient alpha when fit to data from a congeneric population with correlated errors. Given a correct

congeneric sample model, SEM estimates showed little to no bias.

Implications

Applied to a practical setting, the findings of these simulations have important implications for researchers estimating reliability with coefficient omega. The high bias across all one-factor sample models fit to complex populations highlights the importance of specifying multidimensionality. Even in the best of conditions, the simple one-factor sample model overestimated omega by a minimum of .07 when specified to populations with error correlations or multidimensionality. Assuming a one-factor model without specifying error correlations – even for a unidimensional measure – is therefore inappropriate under any circumstance. However, modeling some complexity reduced the average bias of omega, even when the sample model was still incorrect. This is best exemplified by the results for the bifactor population model; specifically for the 16-item low reliability conditions. The simple one-factor sample model overestimated reliability by an average of .28, but specifying correlated errors related to specific factors reduced mean bias to .07, even when sample sizes were small. Omega-hierarchical, however, showed no bias on average, and good precision given a reasonable sample size for this condition. This suggests that as the modeling of multidimensionality approaches the correct model, omega estimates became less biased. Additionally, although omega-hierarchical in particular was relatively robust against model misspecification, the most accurate estimates most consistently resulted from correctly specified models. Choosing an appropriate CFA model is crucial to producing accurate estimates of reliability with coefficient omega.

CFA models are generally chosen through a combination of theory and evidence, relying on methods such as goodness of fit to help determine the appropriateness of a model for a given measure. Although an overall association was found between omega estimate bias and goodness

of fit, this association broke down when the data were separated by population and sample model. For correctly specified models, the relationship was quite weak, and when the statistics showed good or reasonable fit, bias did not differ by more than .01 or .02 on average. Variability in estimates for as CFI and TLI approached one and RMSEA approached zero was high among incorrectly specified models. Additionally, even when the simple model showed good fit, mean bias remained high. There was also very high variability in bias, particularly among better values of CFI, TLI, and RMSEA, indicating that reliance on goodness of fit statistics is not sufficient to ensure unbiased reliability estimates. Researchers should employ other methods, including theory, previous studies, and other model selection criteria such as AIC and Bayes factors to confirm the best model has been chosen.

Correct model selection nullified the effects of other study factors on mean absolute bias; however, since the true model cannot be known, researchers should still be cautious of interpreting coefficient omega. The average bias for estimates based on misspecified models was higher when true reliability was low and worsened by an increase in number of items. Like true models, true reliability cannot be known and is difficult to manipulate. Researchers should be aware that lower values of omega are more likely to be biased, and that this bias is generally an overestimation, rather than underestimation of the reliability of a given measure. Comparing previous estimates with one's own estimate when possible may also aid in identifying situations where omega has been biased. Researchers should also be aware that when true reliability is low, variability in omega estimates is higher, and that steps can be taken to ensure better precision.

In addition to low population reliability, estimate variability was primarily related to sample size. As with any parametric statistic, larger samples produced lower variability in sample estimates across replications. Larger sample sizes are therefore recommended for more

stable estimates of omega. Although a sample size of 1,000 may be ideal, it is often impractical in applied research. Samples of 250 or higher generally showed a variability close to .02 to .05 among well-specified sample models and is recommended as a baseline sample size for adequate reliability estimation. Sample sizes of 100 resulted in a much higher variability, ranging between .05 and .09, and are unsuitable for stable estimation of omega. Small sample sizes are also unsuitable for factor analysis and for model comparison. As such, researchers who wish to investigate scale reliability should aim for sample sizes compatible with running a good confirmatory factor analysis.

Limitations and Directions for Future Research

This study investigated the potential for bias of two omega estimates under a variety of conditions; however, there are several limitations to its scope which should be addressed by future studies. First, a few cells in this study are based on far fewer than 1,000 replications. For the bifactor population, some of the cells had sample sizes as low as 300 to 500 replications, although only when the sample size was $N = 100$. As such, there is a higher risk of error within the results of these cells.

These simulations also treated true reliability and scale length as binary, and therefore makes assumptions about the pattern of results based on values studied here. While it appears to be the case that as true reliability increases, bias decreases, the amount to which a true population value between .60 and .85 or lower than .60 may bias omega remains uncertain, and the point at which lower reliability becomes problematic has not been addressed. Similarly, model misspecification was constrained to selection of an incorrect model type. The effect of misspecifying items to incorrect factors or failing to model all correlated errors, for example, was not addressed. Further studies should verify these findings for various population parameters and

investigate the effects of minor misspecifications.

Higher-order models were also included at the population level, but not the sample-level, and the performance of omega higher-order (see Flora, 2020) when models are misspecified was not evaluated. Data was also drawn from a multivariate normal distribution, assuming continuous item responses. This was done to examine the effects of model misspecification without the confound of categorical versus continuous responses. In practice, many psychometric measures use ordered, categorical items which require polychoric correlations to estimate an accurate CFA model. In these situations, coefficient omega must be adapted to become omega-categorical (Green & Yang, 2009). Important expansions on these findings will therefore be to investigate the performances of both categorical omega and omega higher-order under varying conditions of model misspecification to confirm whether the above guidelines can be generalized.

Finally, these simulations were based on confirmatory factor analysis. CFA requires strong theory or evidence and is used to verify the hypothesized factor structure of a given measure. Particularly in the scale development phase, researchers may wish to estimate reliability for measures using exploratory factor analysis (EFA) when the factor structure is uncertain. EFA models are estimated using different methods from CFA, and often require the use of rotation techniques which alter factor loadings for interpretability. As such, omega estimates must be adapted to account for these changes, and exploratory omega is better suited to estimate corresponding reliability (Flora, 2020). Future studies should further investigate the performance of exploratory omega.

Conclusion

A series of Monte Carlo simulations indicated that coefficient omega is a relatively unbiased estimate of reliability when the sample model is correctly specified. Unidimensional omega based on a sample model with no correlated errors was appropriate only when correctly specified, as it was highly biased when the true model was more complex. In cases of misspecification of multidimensional models, unidimensional omega performed well only when correlated errors pertaining to specific or lower-order factors were included in the model. Bias was higher when tests had a lower true reliability and more items. Omega-hierarchical, based on a bifactor sample model, was the most robust to model misspecification, showing an average bias of +/- .02 when fit to a one-factor model with correlated errors or a higher-order model. The best performance for all populations were observed when models were correctly specified, as long as sample size was sufficient ($N \geq 250$). Researchers who wish to estimate reliability using omega should therefore take care to find the most accurate CFA model to fit their scale, based not only on fit indices, but also on theory and previous evidence, and ensure a reasonable sample size before running their analyses.

References

Black, R. A., Yang, Y., Beitra, D., & McCaffrey, S. (2015). Comparing fit and reliability estimates of a psychological instrument using second-order CFA, bifactor, and essentially tau-equivalent (coefficient alpha) models via AMOS 22. *Journal of Psychoeducational Assessment, 33*(5), 451-472. http://dx.doi.org/10.1177/0734282914553551.

Bonifay, W. & Cai, L. (2017). On the complexity of item response theory model. *Multivariate Behavioral Research, 52*(4), 465-484. http://dx.doi.org/10.1080/00273171.2017.1309262.

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5*(1), 184-186. http://dx.doi.org/10.1177/2167702616657069.

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

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399-412. http://dx.doi.org/10.1111/bjop.12046.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370-378. http://dx.doi.org/10.1177/1948550617693063.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 484-501.

http://dx.doi.org/10.1177/2515245920951747

Green, S. B. & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155-167. http://dx.doi.org/10.1007/s11336-008-90099-3.

Green, S. B. & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement, 34*(4), 14-20. http://dx.doi.org/10.1111/emip.12100.

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. *Methodology, 9*(1), 30-40. http://dx.doi.org/10.1027/1614-2241/a000052.

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

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). semTools: Useful tools for structural equation modeling. R package. http://CRAN.R-project.org/package=semTools.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods, 21*(1), 69–92. http://doi.org/10.1037/a0040086.

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

Maxwell, A. E. (1968). The effect of correlated errors on estimates of reliability coefficients. *Educational and Psychological Measurement, 28*, 803-811. http://dx.doi.org/10.1177/001316446802800309.

Maydeu-Olivares, A. & Coffman, D. L. (2006). Random intercept factor analysis. *Psychological Methods, 11*(4), 344-362. http://dx.doi.org/10.1037/1082-989X.11.4.344.

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

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence, 3*, 2-20. http://dx.doi.org/10.3390/jintelligence3010002.

Murray, A. J. & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Journal of Intelligence, 41*, 407-422. http://dx.doi.org/10.1016/j.intell.201306.004.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

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

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69-76. http://dx.doi.org/10.1177/01466216010251005.

Raykov, T., & Marcoulides, G. A. (2011). Introduction to psychometric theory. Routledge.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696. http://dx.doi.org/10.1080/00273171.2012.715555.

Reise, S. P., Kim, D. S., Mansolf, M. & Widaman, K. F. (2016). Is the bifactor model a better

model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behaviooral Research, 51*(6), 818-838 http://dx.doi.org/10.1080/00273171.2016.1243461.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*, 51-67. http://dx.doi.org/10.1207/s15327906mbr2301_3.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. http://dx.doi.org/10.1037/met0000045.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36.

Savalei, V. & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology*, *5*(1), 36. http://dx.doi.org/10.1525/collabra.247.

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

Stanley, L. M. & Edwards, M. C. (2016). Reliability and model fit. *Educational Psychological Measurement, 76*(6), 976-985. http://dx.doi.org/10.1177/001364416638900.

Trizano-Hermosilla, I. & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology, 7*. http://dx.doi.org/10.3389/fpsyg.2016.00769.

Venables, W. N. & Ripley, B. D. (2002) Modern applied statistics with S (4th ed.). Springer: New York.

Yang, Y. & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17*, 66-81.

Yang, Y. & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377-392. http://dx.doi.org/10.1177/0734282911406668.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. http://dx.doi.org/10.1007/s11336-003-0974-7.

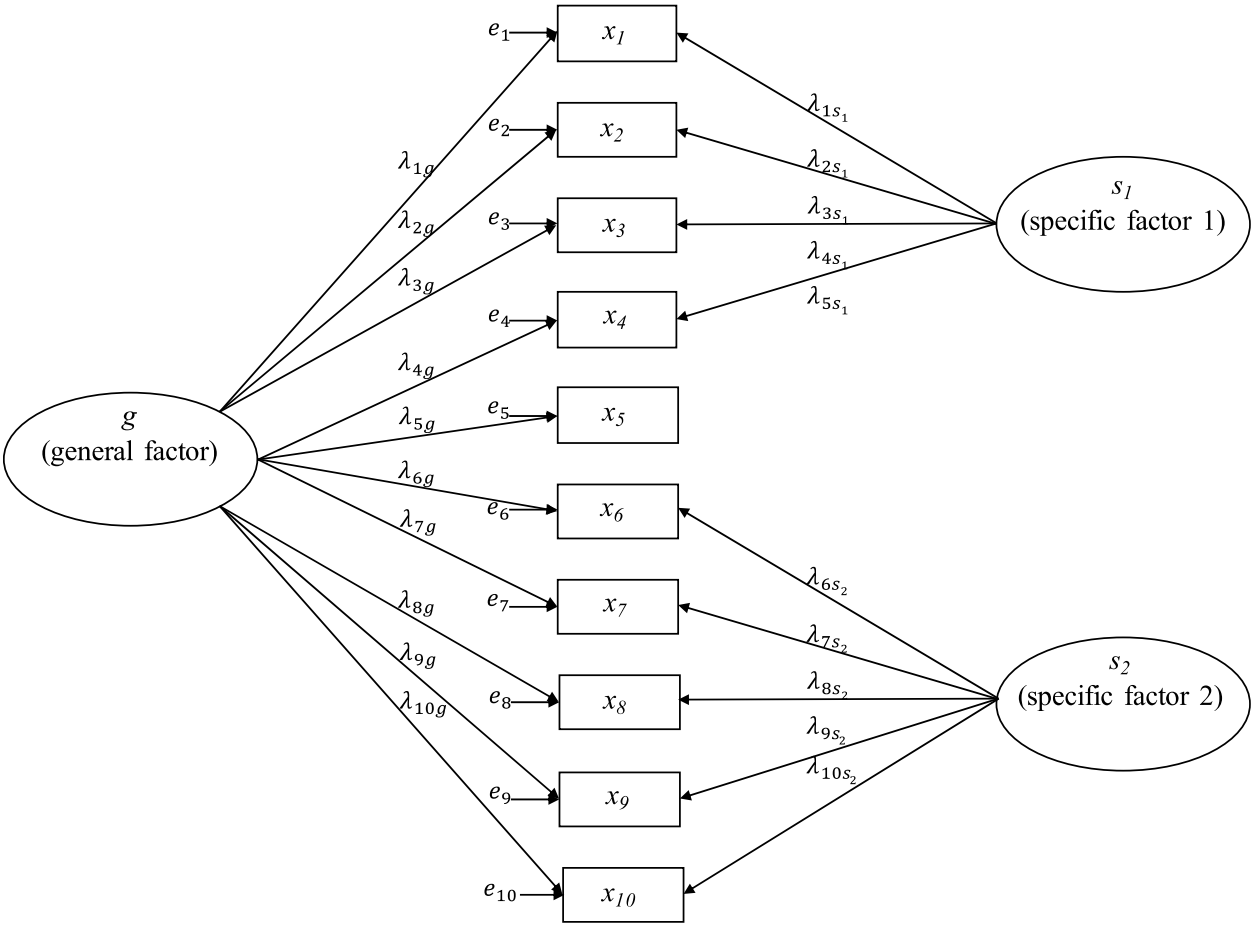

Figure 1. Example path diagram for a fictional bifactor scale comprised of ten items. Observed scores are represented by x_j, where j is the item number. All items relate to the general factor (g) with strength quantified by its factor loading (λ_{jg}). The first five items are also related to a specific factor (s_1) while the second five items are related to a second specific factor (s_2), capturing residual covariance for each subset of items that remains over and above the general factor. Each item has its own unique error term (e_j).

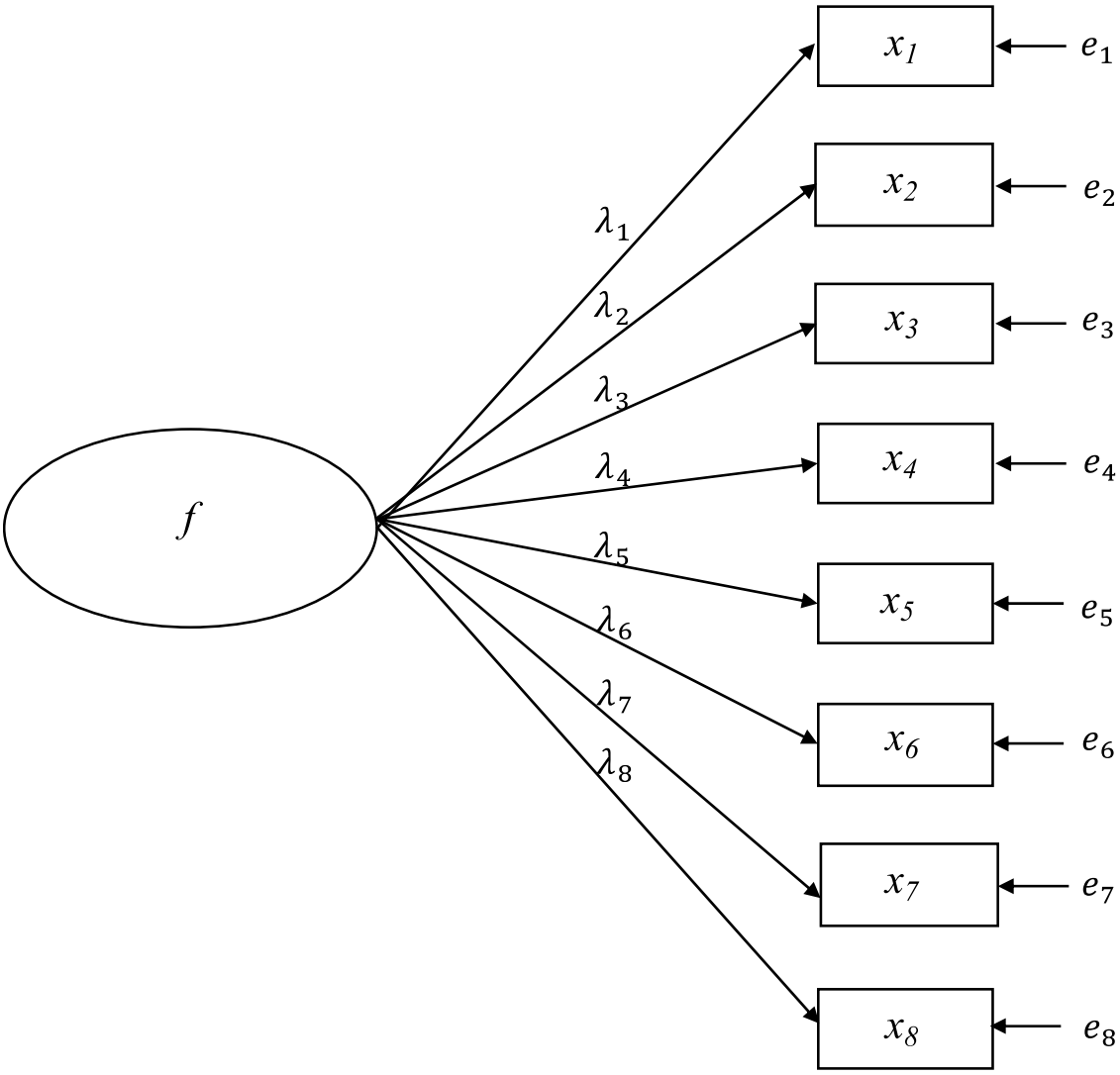

Figure 2. Simple one-factor population model from which samples will be drawn. Items load with a strength of λ_j onto a single factor. Errors are uncorrelated and unique to each item.

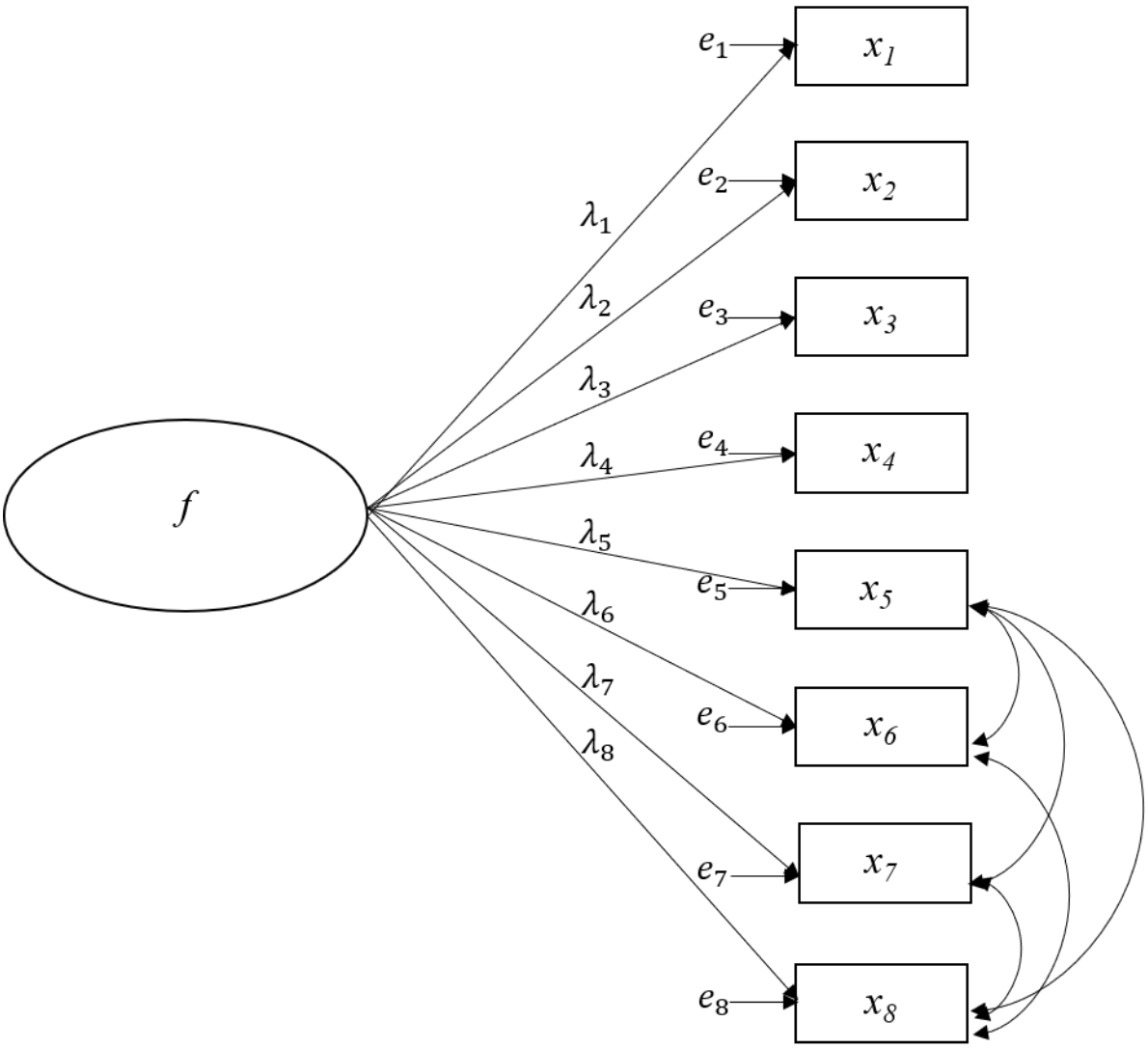

Figure 3. One-factor population model with correlated errors from which samples will be drawn. Items load with a strength of λ_j onto a single factor. Errors are correlated among the second half of the items.

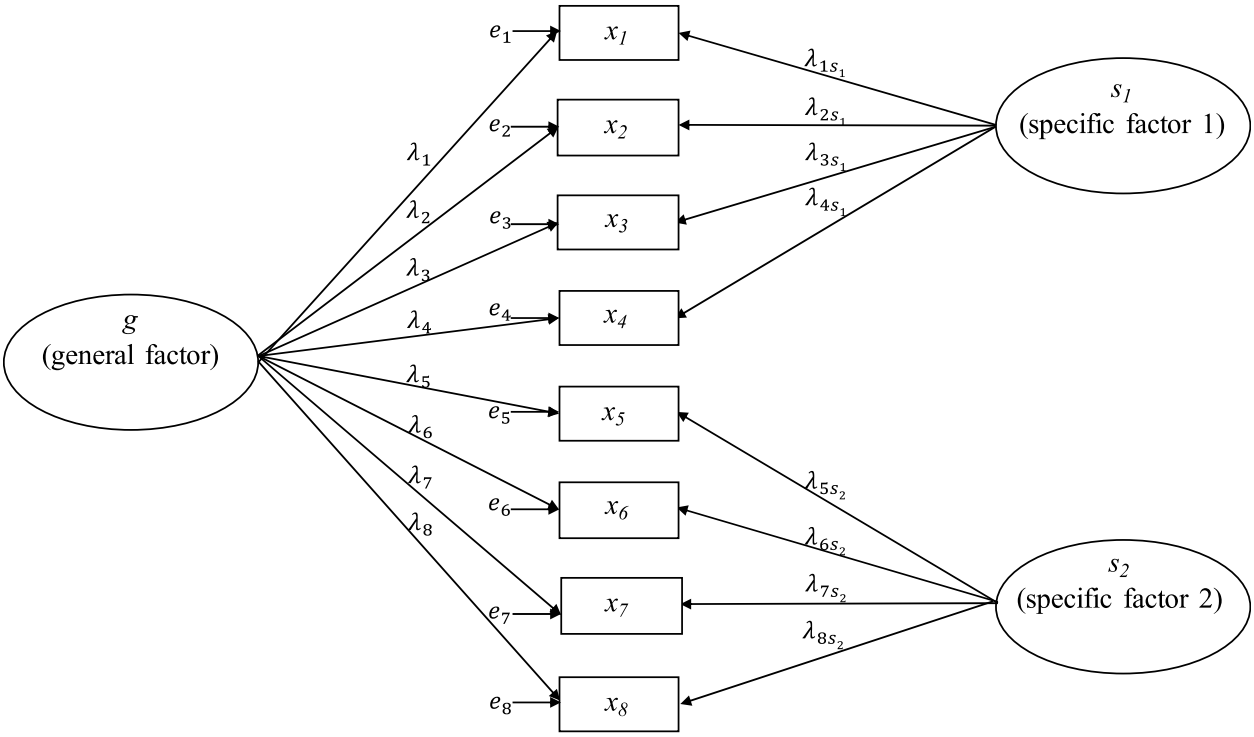

Figure 4. Bifactor population model from which samples will be drawn. Items load with a strength of λ_j onto a single general factor, while two specific factors capture residual covariance among subsets of items. Errors are uncorrelated and unique to each item.

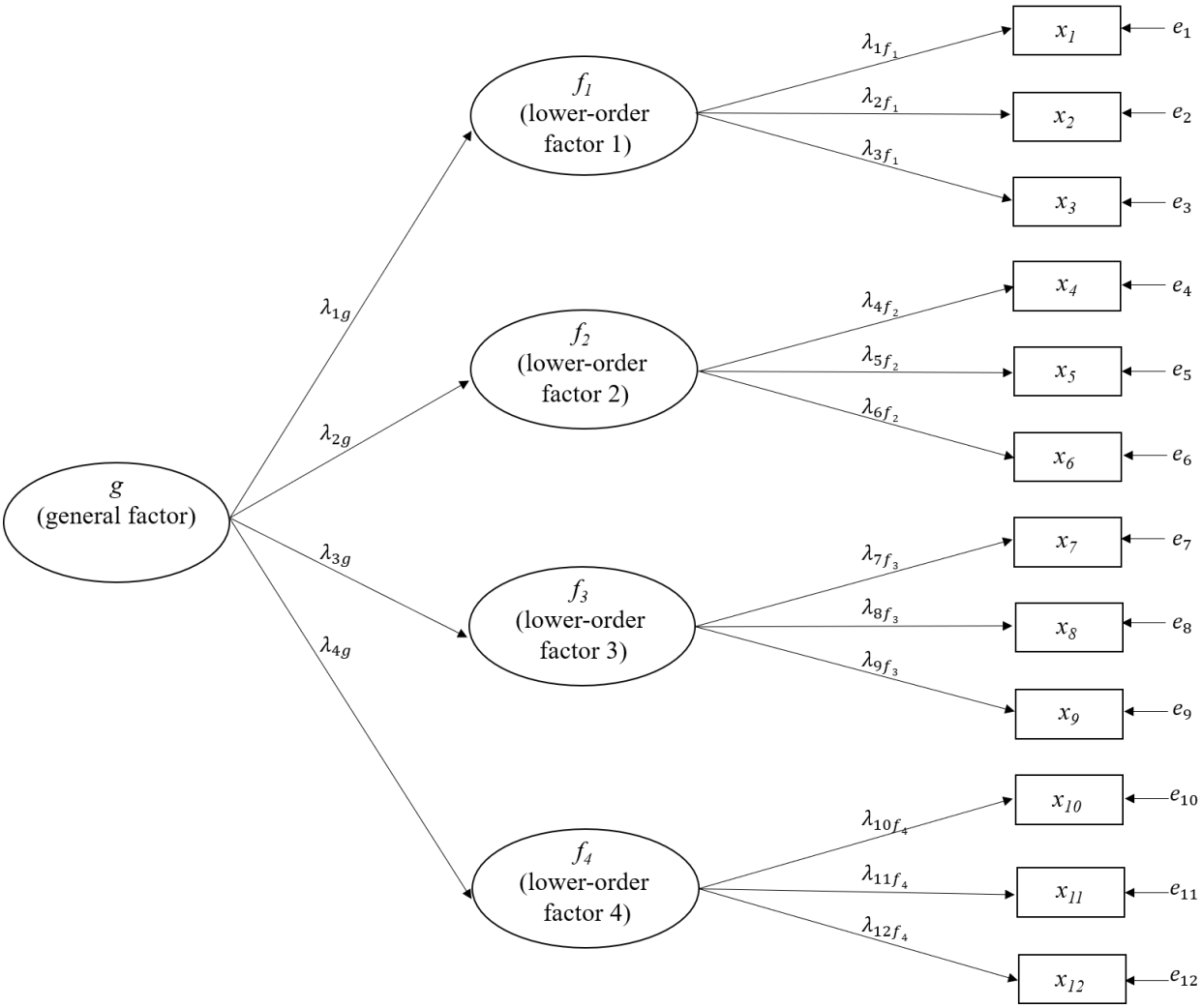

Figure 5. Higher order population model from which samples will be drawn. A single general factor influences items through four lower-order factors. Errors are uncorrelated and unique to each item.

Table 1
Population factor loadings and error covariances for each true model.

Model	Factor loading matrix dimension	Population factor loadings	Error covariances
Simple one-factor, 8 items, low reliability	8x1	.414, .210, .472, .416, .325, .504, .301, .521	(none)
Simple one factor, 8 items, high reliability	8x1	.847, .423, .870, .516, .648, .721, .322, .743	(none)
Simple one-factor, 16 items, low reliability	16x1	.245, .423, .229, .316, .420, .414, .224, .403, .312, .248, .331, .266, .391, .202, .165, .104	(none)
Simple one factor, 16 items, high reliability	16x1	.745, .423, .730, .416, .628, .514, .324, .643, .612, .548, .331, .246, .741, .502, .365, .316	(none)
Correlated errors, 8 items, low reliability	8x1	.504, .293, .412, .506, .451, .574, .434, .610	.287, .343, .334, .232 .197, .312
Correlated errors, 8 items, high reliability	8x1	.823, .593, .837, .706, .751, .834, .746, .744	.187, .263, .264, .232, .197, .282
Correlated errors, 16 items, low reliability	16x1	.504, .293, .412, .396, .401, .574, .414, .410, .471, .532, .624, .523, .339, .419, .331, .345	.347, .413, .294, .418, .524, .359, .236, .232, .217, .214, .227, .335, .304, .322, .344, .220, .381, .392, .288, .271, .401, .310, .447, .313, 524, .232, .357, 392
Correlated errors, 16 items, high reliability	16x1	.504, .293, .412, .396, .401, .574, .414, .410, .471, .532, .624, .523, .339, .419, .331, .345	.347, .413, .294, .418, .524, .359, .236, .232, .217, .214, .227, .335, .304, .322, .344, .220, .381, .392, 288, 271, 401, .310, .447, .313, .524, .232, .357, .392
Bifactor, 8 items, low reliability	8x3	.462, .340, .420, .659, .314, .501, .608, .410, .426, .414, .279, .338, 0, 0, 0, 0, 0, 0, 0, 0, .314, .448, .213, .417	(none)
Bifactor, 8 items, high reliability	8x3	.852, .736, .868, .612, .913, .704, .719, .611, .426, .414, .279, .338, 0, 0, 0, 0, 0, 0, 0, 0, .314, .448, .213, .417	(none)
Bifactor, 16 items, low reliability	16x3	.432, .536, .711, .312, .313, .501, .415, .631,	(none)

		.653, .467, .358, .233, .448, .535, .317, .502, .426, .474, .376, .598, .451, .526, .663, .424, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, .519, .448, .414, .618, .269, .367, .389, .420	
Bifactor, 16 items, high reliability	16x3	.852, .636, .868, .612, .513, .704, .719, .831, .653, .774, .358, .721, .448, .835, .612, .718, .426, .374, .276, .238, .351, .248, .319, .314, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, .317, .448, .314, .418, .269, .227, .389, .420	(none)
Higher-order, 12 items, low reliability	12x4x1	.60, .72, .64, .62, .68, .73, .83, .54, .64, .44, .75, .66, .59, .62, .69, .61	(none)
Higher-order, 12 items, high reliability	12x4x1	.84, .87, .89, .82, 84, .86, .83, .91, .89, .84, .85, .84, .79, .72, .91, .81	(none)
Higher-order, 20 items, low reliability	20x4x1	.61, .53, .79, .49, .71, .71, .73, .76, .65, .68, .58, .69, .49, .72, .65, .66, .74, .58, .54, .61, .59, .64, .51, .62	(none)
Higher-order, 20 items, high reliability	20x4x1s	.84, .87, .89, .82, .84, .86, .83, .91, .86, .84, .85, .84, .76, .81, .68, .87, .74, .80, .54, .88, .79, .72, .89, .82	(none)

Table 2
Frequency of converged and proper solutions across cells.

Sample model	Correlated Errors Population				Bifactor Population				Higher-Order Population			
	Correlated Errors		Bifactor		Correlated		Bifactor		Correlated Errors		Bifactor	
	Con.	Pr.	Con.	Pr.	Con.	Pr.	Con.	Pr.	Con.	Pr.	Con.	Pr.
8 items												
High reliability												
n = 100	1000	1000	1000	1000	1000	1000	863	527	1000	998	988	886
n = 250	1000	1000	1000	1000	1000	1000	951	798	1000	1000	1000	990
n = 1000	1000	1000	1000	1000	1000	1000	990	976	1000	1000	1000	1000
Low reliability												
n = 100	999	983	960	872	1000	947	751	323	1000	967	973	549
n = 250	1000	1000	992	983	1000	999	782	504	1000	1000	995	915
n = 1000	1000	1000	1000	1000	1000	1000	915	825	1000	1000	1000	1000
16 items												
High reliability												
n = 100	1000	1000	638	638	1000	993	854	644	1000	1000	989	918
n = 250	1000	1000	829	829	1000	1000	960	821	1000	1000	999	997
n = 1000	1000	1000	894	894	1000	1000	1000	990	1000	1000	1000	1000
Low reliability												
n = 100	1000	999	766	683	1000	333	899	727	1000	891	965	746
n = 250	1000	1000	816	800	1000	483	980	928	1000	999	1000	987
n = 1000	1000	1000	942	941	1000	658	1000	1000	1000	1000	1000	1000

Note: 'Con.' represents the number of models which converged. 'Pr.' represents the number of converged models that produced proper solutions. All models in the simple one-factor population model converged, and all but two produced proper solutions. All simple models fit to data from more complex population models also converged and produced proper solutions, such that N_{rep} = 1000 for all cells.

Table 3
Absolute bias of coefficient omega for samples drawn from a simple one-factor population model.

Condition	ω_{Σ} bias	ω_{S} bias
8 items		
n = 100		
High reliability	.00 (.02)	.00 (.02)
Low reliability	.00 (.07)	-.01 (.07)
n = 250		
High reliability	.00 (.01)	.00 (.01)
Low reliability	.00 (.04)	.00 (.04)
n = 1000		
High reliability	.00 (.01)	.00 (.01)
Low reliability	.00 (.02)	.00 (.02)
16 items		
n = 100		
High reliability	.00 (.02)	.00 (.02)
Low reliability	-.01 (.07)	-.01 (.07)
n = 250		
High reliability	.00 (.01)	.00 (.01)
Low reliability	.00 (.04)	.00 (.04)
n = 1000		
High reliability	.00 (.01)	.00 (.01)
Low reliability	.00 (.02)	.00 (.02)

Note: Absolute bias is presented as the mean bias across 1000 replications for all conditions, except the low reliability condition with 8 items and sample size of 100, for which only 998 replications produced proper solutions. ω_{Σ} represents coefficient omega calculated using the model-implied total variance as the equation denominator. ω_{S} represents coefficient omega calculated using the observed total variance as the equation denominator. High reliability conditions had a population reliability of approximately ρ = .85, while low reliability conditions had a population reliability of approximately ρ = .60.

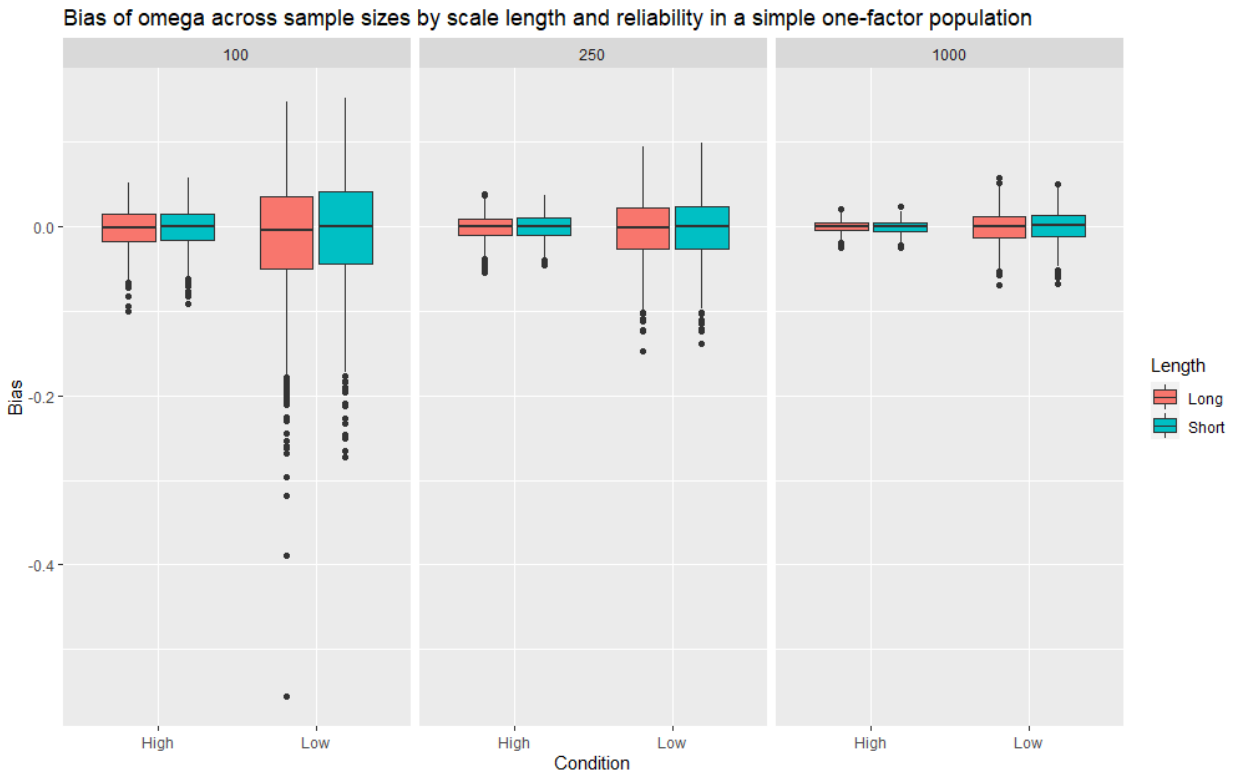

Figure 6. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for sample sizes of 100, 250, and 1000 for an underlying population model that has one factor with correlated errors. For the high reliability conditions, population reliability was approximately ρ = .85, and low reliability conditions had a population reliability of approximately ρ = .60. Length refers to the number of items, such that long tests have 16 items and short tests have 8 items.

Table 4

Absolute bias of coefficient omega for samples drawn from a population with one-factor with correlated errors.

	Simple			Correlated Errors			Bifactor		
	N	ω_Σ bias	ω_S bias	*N*	ω_Σ bias	ω_S bias	*N*	ω_Σ bias	ω_S bias
8 items									
n = 100									
High reliability	1000	.07 (.01)	.07 (.01)	1000	**.00 (.03)**	.00 (.03)	1000	.00 (.03)	.00 (.03)
Low reliability	1000	.17 (.04)	.15 (.04)	983	**.00 (.09)**	.00 (.09)	872	-.01 (.09)	-.01 (.09)
n = 250									
High reliability	1000	.07 (.01)	.06 (.01)	1000	**.00 (.02)**	.00 (.02)	1000	.00 (.02)	.00 (.02)
Low reliability	1000	.17 (.02)	.15 (.03)	1000	**.00 (.06)**	.00 (.06)	983	.00 (.06)	.00 (.06)
n = 1000									
High reliability	1000	.08 (.00)	.06 (.01)	1000	**.00 (.01)**	.00 (.01)	1000	.00 (.01)	.00 (.01)
Low reliability	1000	.17 (.01)	.15 (.01)	1000	**.00 (.03)**	.00 (.03)	1000	.00 (.03)	.00 (.03)
16 items									
n = 100									
High reliability	1000	.09 (.01)	.07 (.01)	1000	**.00 (.03)**	.00 (.03)	638	.00 (.03)	.00 (.03)
Low reliability	1000	.25 (.02)	.22 (.03)	999	**.00 (.08)**	.00 (.08)	683	.02 (.10)	.01 (.09)
n = 250									
High reliability	1000	.09 (.01)	.09 (.01)	1000	**.00 (.02)**	.00 (.02)	829	.00 (.02)	.00 (.02)
Low reliability	1000	.25 (.01)	.22 (.02)	1000	**.00 (.05)**	.00 (.05)	800	.02 (.06)	.01 (.06)
n = 1000									
High reliability	1000	.09 (.00)	.07 (.00)	1000	**.00 (.01)**	.00 (.01)	984	.00 (.01)	.00 (.01)
Low reliability	1000	.25 (.01)	.22 (.01)	1000	**. 00 (.02)**	.00 (.02)	941	.01 (.03)	.01 (.03)

Note: Bold values represent the bias for the correctly specified model. Standard deviations of absolute bias are marked in brackets beside mean absolute bias. *N* represents the number of converged, proper solutions, while *n* represents the sample size within each replication. ω_Σ represents coefficient omega calculated using the model-implied total variance as the equation denominator. ω_S represents coefficient omega calculated using the observed total variance as the equation denominator.

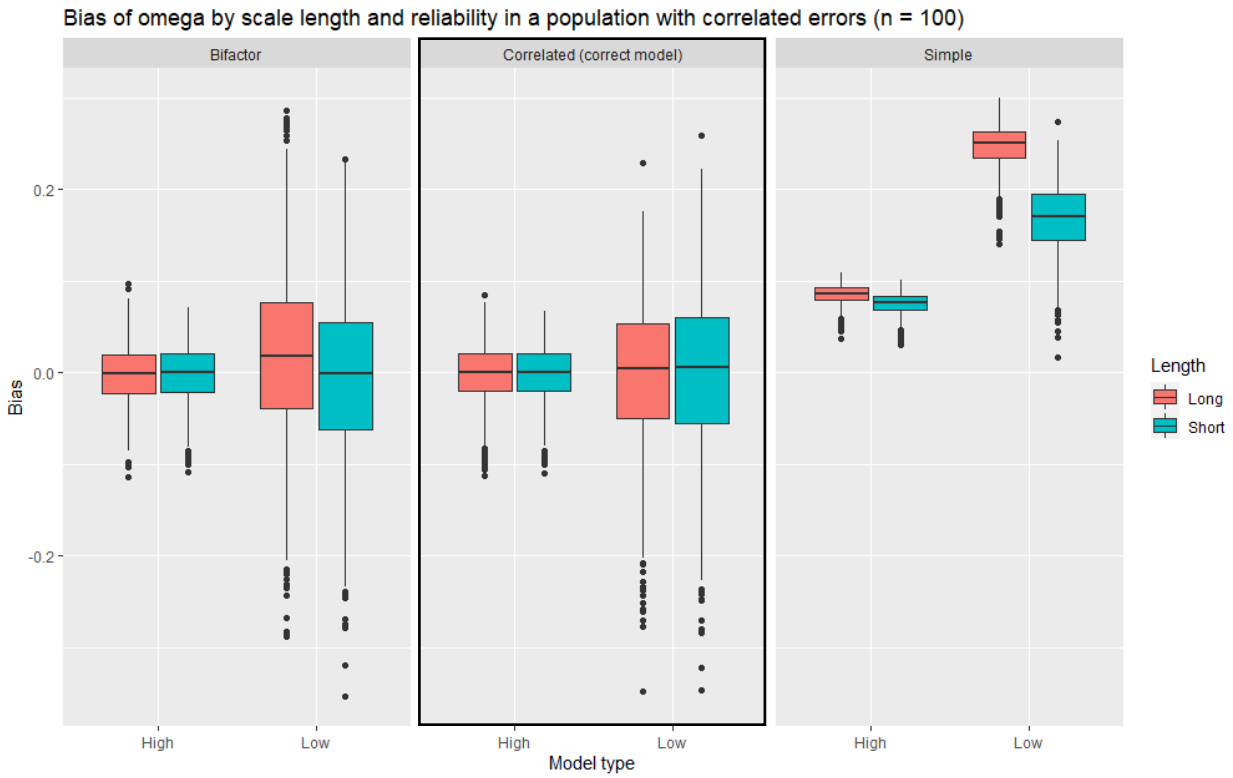

Figure 7. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 100 participants for an underlying population model that has one factor with correlated errors. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

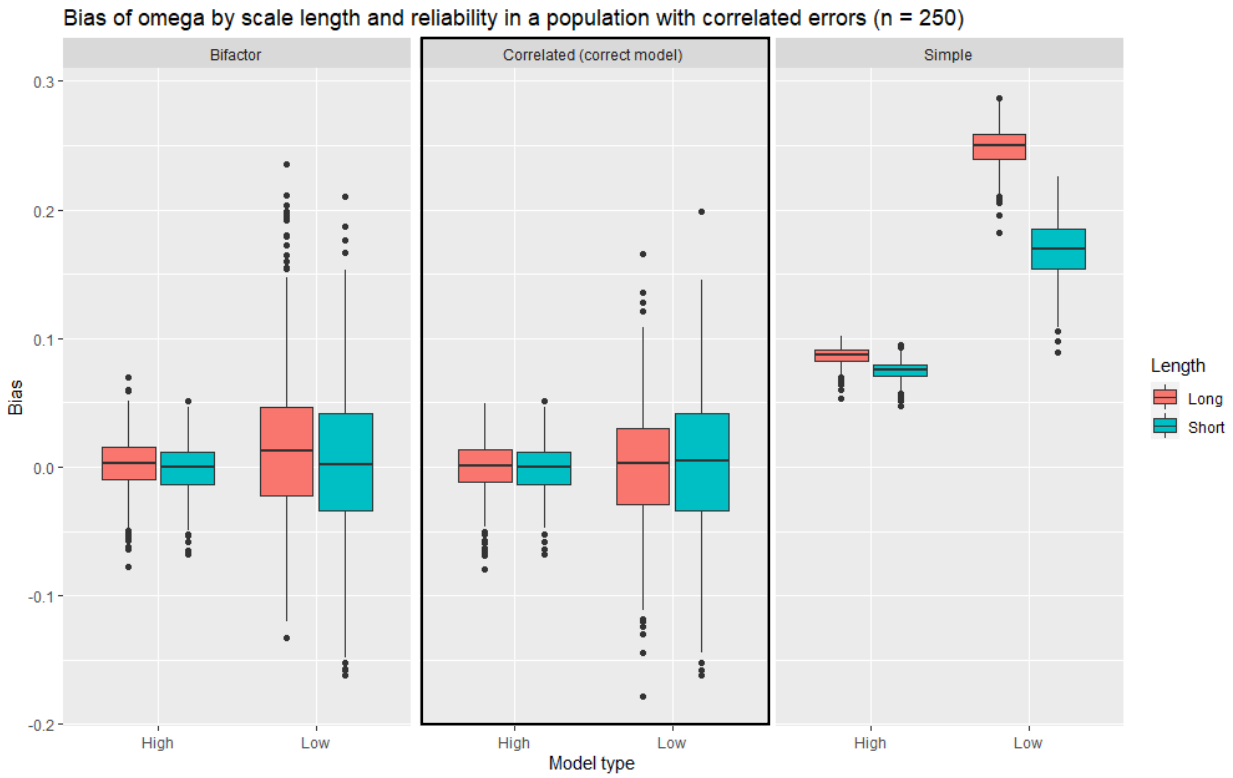

Figure 8. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 250 participants for an underlying population model that has one factor with correlated errors. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

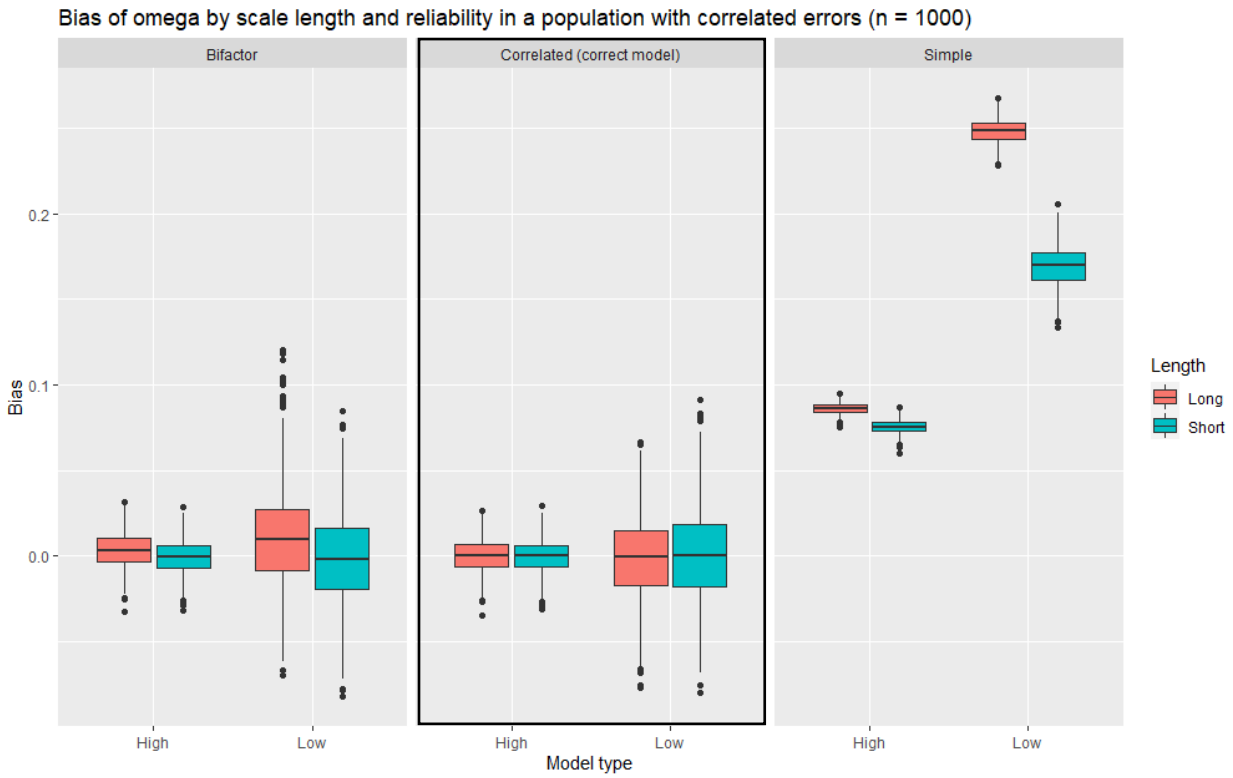

Figure 9. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 1000 participants for an underlying population model that has one factor with correlated errors. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

Table 5

Absolute bias of coefficient omega for samples drawn from a bifactor population.

	Simple			Correlated Errors			Bifactor		
	N	ω_Σ bias	ω_S bias	*N*	ω_Σ bias	ω_S bias	*N*	ω_Σ bias	ω_S bias
8 items									
n = 100									
High reliability	1000	.08 (.01)	.08 (.01)	1000	.00 (.03)	.00 (.03)	527	**.00 (.03)**	.00 (.03)
Low reliability	1000	.14 (.04)	.14 (.05)	947	.00 (.07)	.00 (.08)	323	**.00 (.09)**	.00 (.09)
n = 250									
High reliability	1000	.08 (.01)	.08 (.01)	1000	.00 (.02)	.00 (.02)	798	**.00 (.02)**	.00 (.02)
Low reliability	1000	.15 (.03)	.14 (.03)	999	.01 (.05)	.01 (.05)	504	**.01 (.05)**	.01 (.05)
n = 1000									
High reliability	1000	.08 (.00)	.08 (.00)	1000	.00 (.01)	.00 (.01)	976	**.00 (.01)**	.00 (.01)
Low reliability	1000	.15 (.01)	.15 (.01)	1000	.01 (.02)	.01 (.02)	825	**.01 (.03)**	.01 (.03)
16 items									
n = 100									
High reliability	1000	.09 (.01)	.09 (.01)	993	.00 (.03)	.00 (.03)	644	**.00 (.03)**	.00 (.03)
Low reliability	1000	.28 (.02)	.25 (.04)	333	.07 (.06)	.06 (.07)	727	**.01 (.08)**	.01 (.09)
n = 250									
High reliability	1000	.09 (.01)	.09 (.01)	1000	.00 (.02)	.00 (.02)	821	**.00 (.02)**	.00 (.02)
Low reliability	1000	.28 (.01)	.25 (.02)	483	.06 (.03)	.06 (.04)	928	**.01 (.05)**	.01 (.05)
n = 1000									
High reliability	1000	.10 (.00)	.09 (.00)	1000	.00 (.01)	.00 (.01)	990	**.00 (.01)**	.00 (.01)
Low reliability	1000	.28 (.01)	.26 (.01)	658	.05 (.02)	.05 (.02)	1000	**.00 (.03)**	.00 (.03)

Note: Bold values represent the bias for the correctly specified model. Standard deviations of absolute bias are marked in brackets beside mean absolute bias. *N* represents the number of converged, proper solutions, while *n* represents the sample size within each replication. ω_Σ represents coefficient omega calculated using the model-implied total variance as the equation denominator. ω_S represents coefficient omega calculated using the observed total variance as the equation denominator.

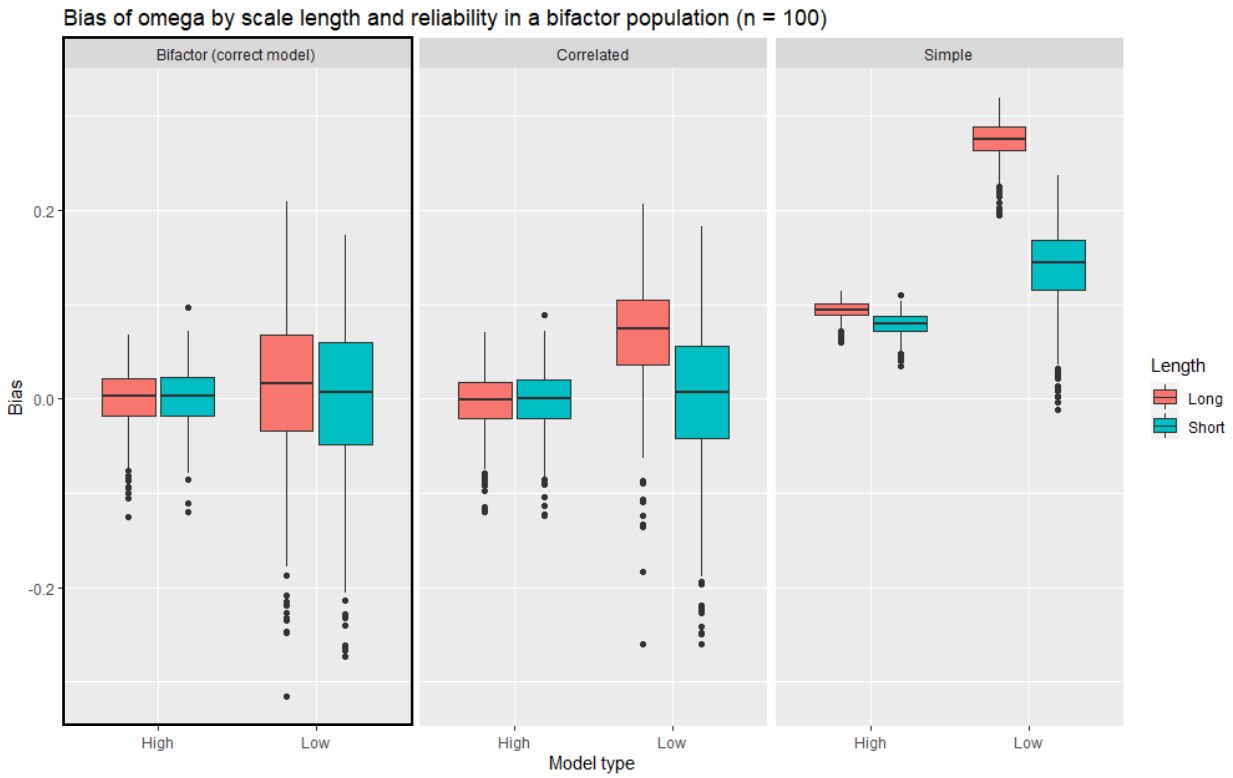

Figure 10. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 100 participants for an underlying population model is bifactor. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

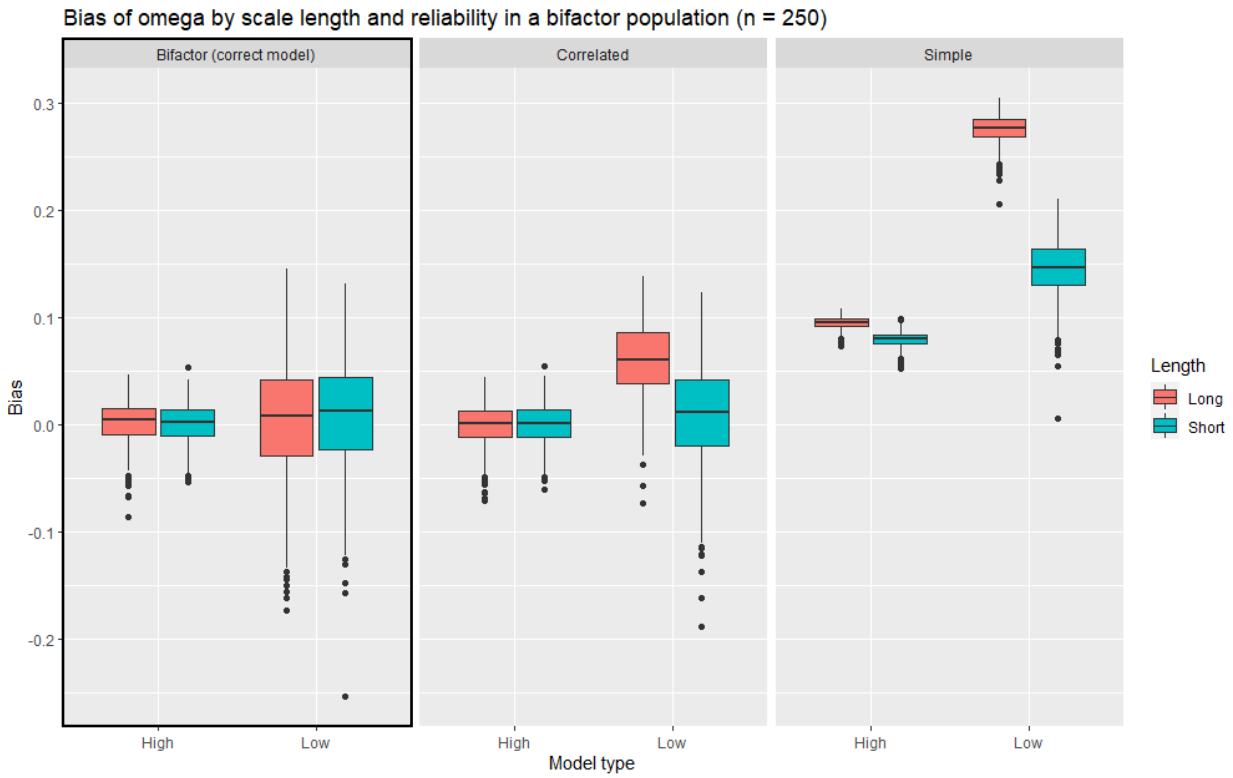

Figure 11. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 250 participants for an underlying population model is bifactor. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

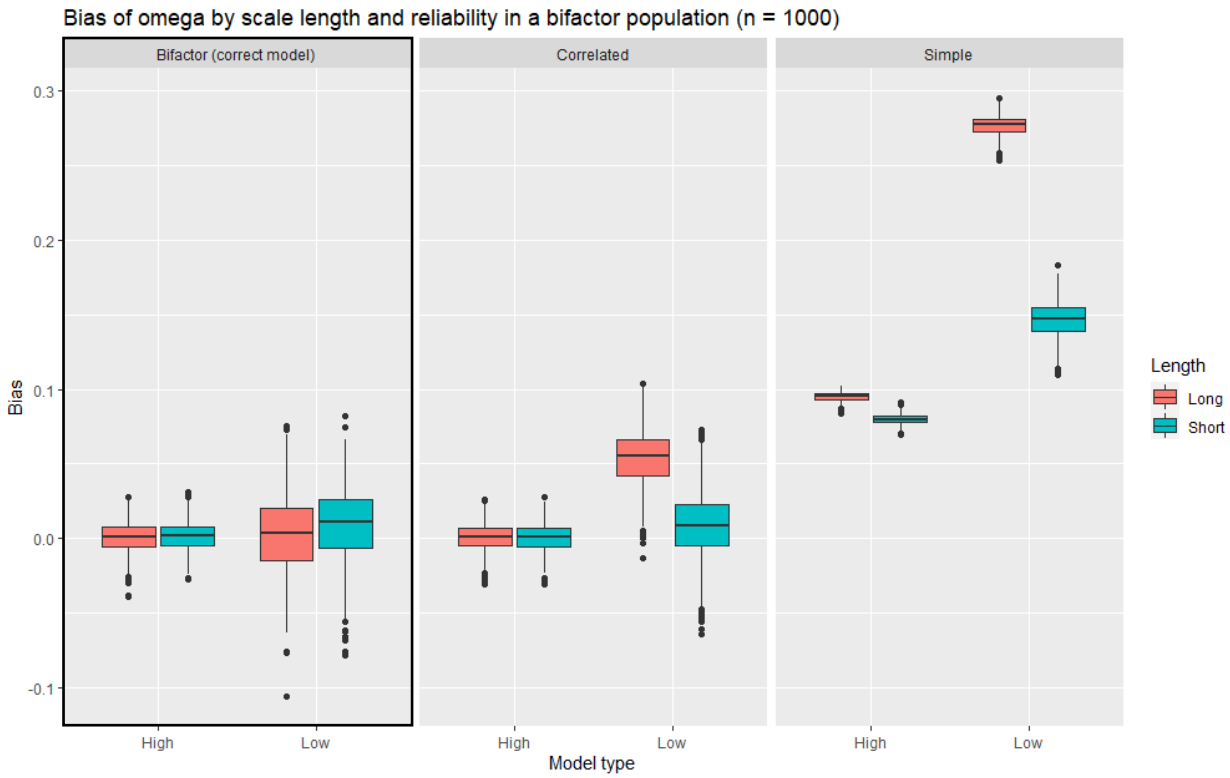

Figure 12. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 1000 participants for an underlying population model is bifactor. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

Table 6
Absolute bias of coefficient omega for samples drawn from a higher-order population model.

	Simple			Correlated Errors			Bifactor		
	N	ω_Σ bias	ω_S bias	*N*	ω_Σ bias	ω_S bias	*N*	ω_Σ bias	ω_S bias
12 items									
n = 100									
High reliability	1000	.08 (.01)	.07 (.01)	998	.00 (.03)	.00 (.03)	886	.00 (.03)	.00 (.03)
Low reliability	1000	.16 (.04)	.14 (.06)	967	.00 (.07)	-.01 (.08)	549	-.01 (.08)	-.02 (.08)
n = 250									
High reliability	1000	.08 (.01)	.07 (.01)	1000	.00 (.02)	.00 (.02)	990	.00 (.02)	.00 (.02)
Low reliability	1000	.16 (.02)	.16 (.03)	1000	.00 (.04)	.00 (.04)	915	.00 (.04)	.00 (.04)
n = 1000									
High reliability	1000	.08 (.00)	.07 (.00)	1000	.00 (.01)	.00 (.01)	1000	.00 (.01)	.00 (.01)
Low reliability	1000	.17 (.01)	.16 (.01)	1000	.00 (.02)	.00 (.02)	1000	.00 (.02)	.00 (.02)
20 items									
n = 100									
High reliability	1000	.10 (.01)	.10 (.01)	1000	.00 (.03)	.00 (.03)	918	.00 (.03)	.00 (.03)
Low reliability	1000	.22 (.03)	.19 (.06)	891	-.02 (.09)	-.04 (.10)	746	-.02 (.08)	-.03 (.09)
n = 250									
High reliability	1000	.10 (.00)	.10 (.01)	1000	.00 (.02)	.00 (.02)	997	.00 (.02)	.00 (.02)
Low reliability	1000	.23 (.02)	.20 (.04)	999	.00 (.04)	-.01 (.04)	987	.00 (.04)	-.01 (.04)
n = 1000									
High reliability	1000	.10 (.00)	.10 (.00)	1000	.00 (.01)	.00 (.01)	1000	.00 (.01)	.00 (.01)
Low reliability	1000	.23 (.01)	.20 (.02)	1000	.00 (.02)	.00 (.02)	1000	.00 (.02)	.00 (.02)

Note: Standard deviations of absolute bias are marked in brackets beside mean absolute bias. *N* represents the number of converged, proper solutions, while *n* represents the sample size within each replication. ω_Σ represents coefficient omega calculated using the model-implied total variance as the equation denominator. ω_S represents coefficient omega calculated using the observed total variance as the equation denominator.

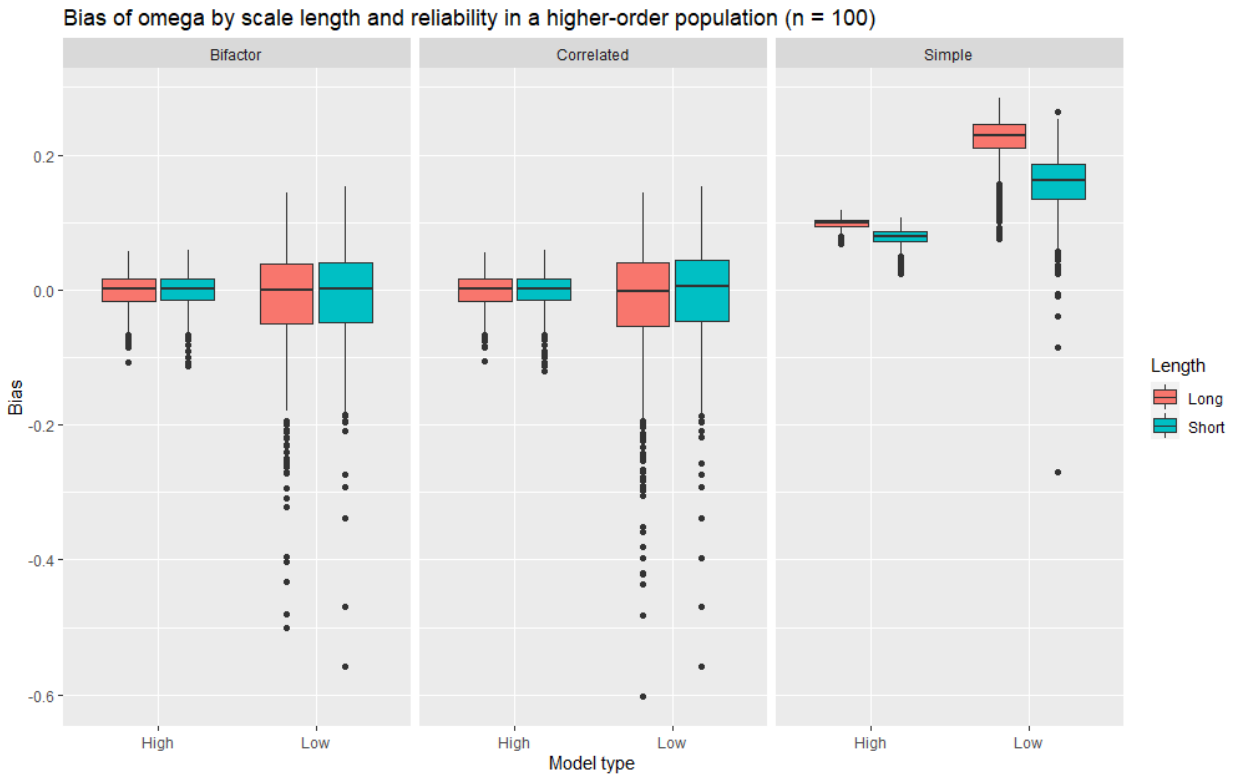

Figure 13. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 100 participants for an underlying population model is higher-order. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

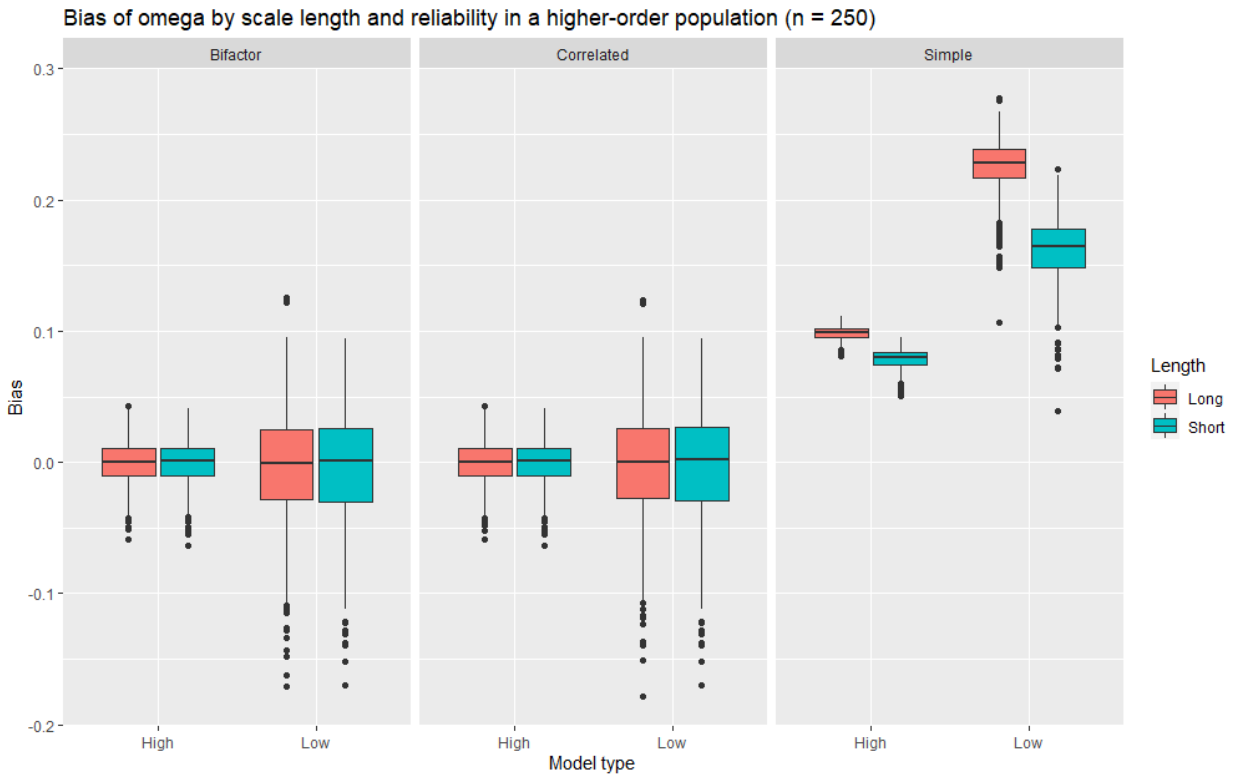

Figure 14. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 250 participants for an underlying population model is higher-order. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

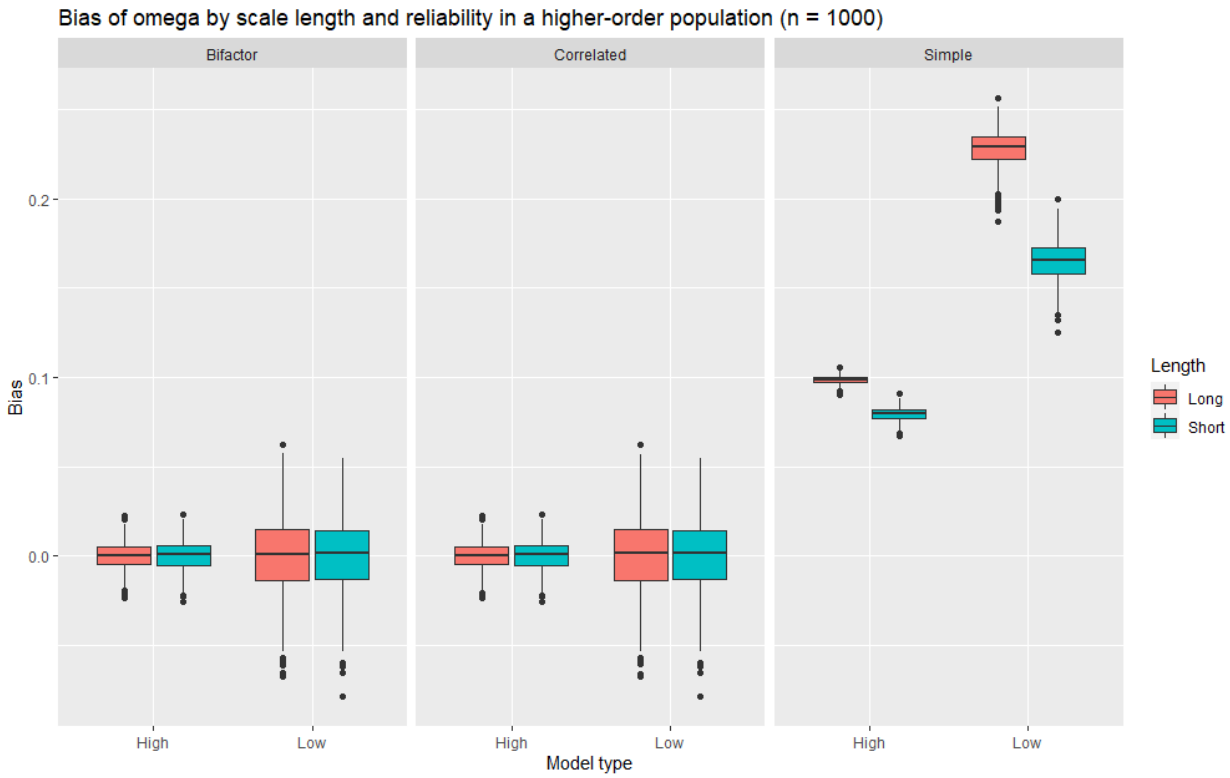

Figure 15. Boxplots showing the absolute bias of coefficient omega using the model-implied total variance as its denominator for a sample of 1000 participants for an underlying population model is higher-order. Bias is shown for three different sample models when the reliability is high (ρ = .85) or low (ρ = .60) and different scale lengths vary such that long tests have 16 items and short tests have 8 items.

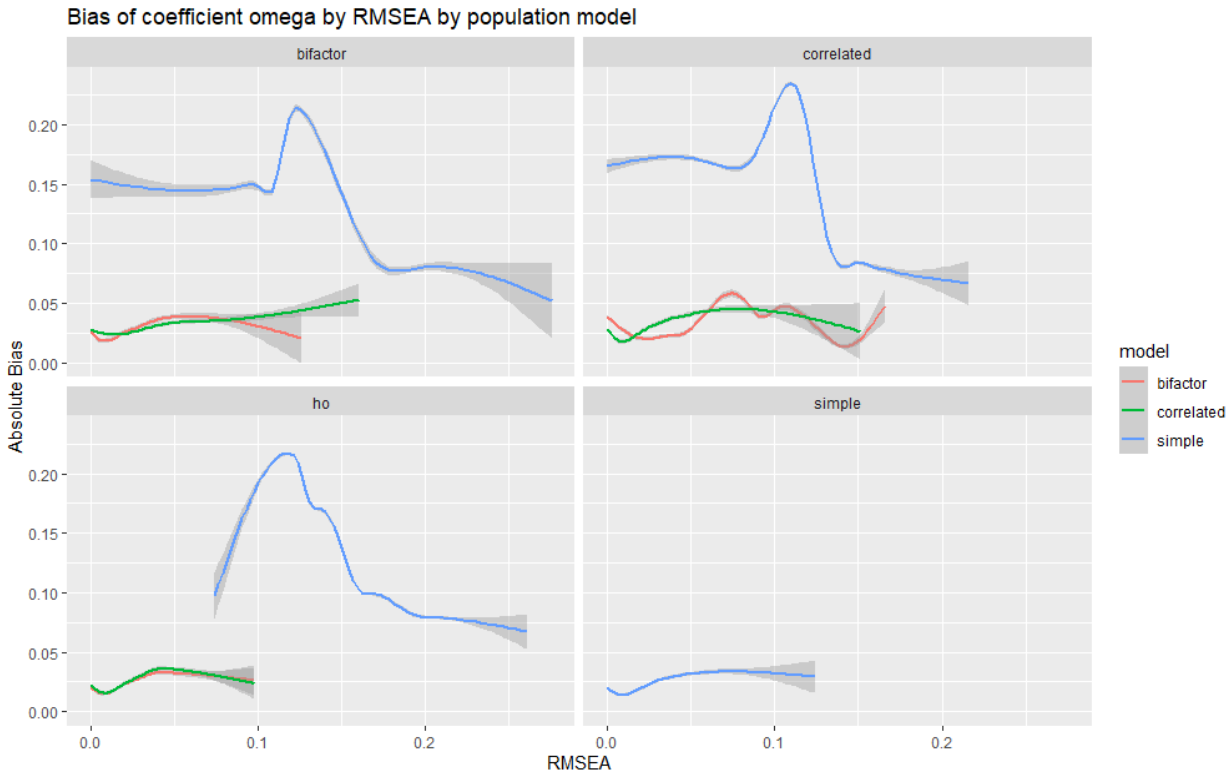

Figure 16. Scatterplot of absolute bias of coefficient omega by RMSEA for each sample model within each population. Absolute bias is represented by the absolute value of bias.

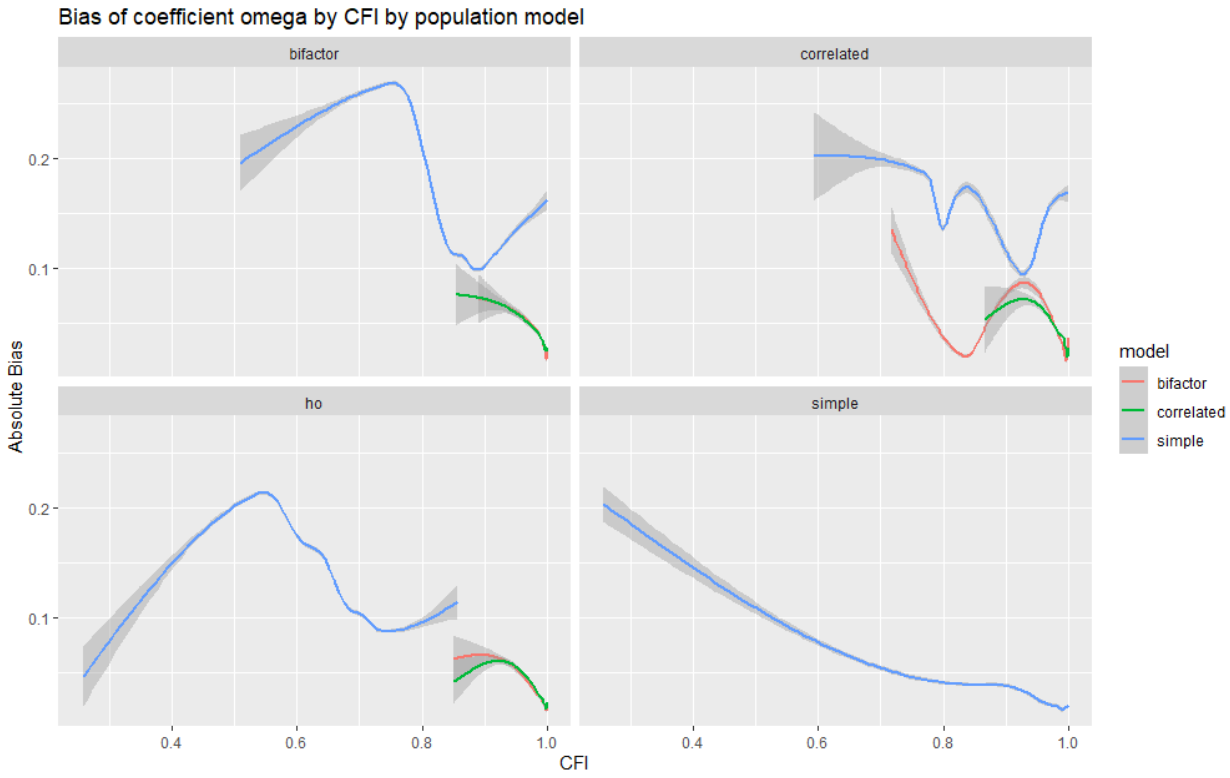

Figure 17. Scatterplot of absolute bias of coefficient omega by CFI for each sample model within each population. Absolute bias is represented by the absolute value of bias.

Figure 18. Scatterplot of absolute bias of coefficient omega by TLI. Absolute bias is represented by the absolute value of bias.

Table 7

Correlations between degree of absolute bias and model fit indices.

	Simple	Correlated Errors				Bifactor				Higher-Order			
Fit statistic	Ov.	Ov.	Sim.	Cor.	Bif.	Ov.	Sim.	Cor.	Bif.	Ov.	Sim.	Cor.	Bif.
RMSEA	.11	.42	-.60	.10	-.10	.64	-.45	.05	.06	.65	-.88	.14	.12
CFI	-.16	-.51	-.24	-.13	.04	-.73	-.50	-.10	-.10	-.74	-.74	-.18	-.17
TLI	-.06	-.48	-.24	-.03	.05	-.68	-.47	-.03	.02	-.71	-.68	-.10	-.10

Note: Correlations represent the Spearman correlation between each respective fit statistic and the absolute bias of coefficient omega estimates using the model-implied total variance denominator. Absolute bias is represented by the absolute value of bias. *Ov*. refers to the overall correlation between the fit statistic and bias across all sample models for a given population model. *Sim*. refers to the simple one-factor sample model. *Cor*. refers to the one-factor sample model with correlated errors. *Bif*. refers to the bifactor sample model.