MULTIVARIATE BEHAVIORAL RESEARCH
2022, VOL. 57, NO. 4, 620–641

Bayesian Estimation of Single-Test Reliability Coefficients

Julius M. Pfadt[a], Don van den Bergh[b], Klaas Sijtsma[c], Morten Moshagen[a], and Eric-Jan Wagenmakers[b]

[a]Department of Psychological Research Methods, Ulm University; [b]Department of Psychological Methods, University of Amsterdam; [c]Department of Methodology and Statistics, Tilburg University

ABSTRACT

Popular measures of reliability for a single-test administration include coefficient α, coefficient λ_2, the greatest lower bound (glb), and coefficient ω. First, we show how these measures can be easily estimated within a Bayesian framework. Specifically, the posterior distribution for these measures can be obtained through Gibbs sampling – for coefficients α, λ_2, and the glb one can sample the covariance matrix from an inverse Wishart distribution; for coefficient ω one samples the conditional posterior distributions from a single-factor CFA-model. Simulations show that – under relatively uninformative priors – the 95% Bayesian credible intervals are highly similar to the 95% frequentist bootstrap confidence intervals. In addition, the posterior distribution can be used to address practically relevant questions, such as "what is the probability that the reliability of this test is between .70 and .90?", or, "how likely is it that the reliability of this test is higher than .80?" In general, the use of a posterior distribution highlights the inherent uncertainty with respect to the estimation of reliability measures.

KEYWORDS

Bayesian reliability estimation; Cronbach's alpha; Guttman's lambda-2; greatest lower bound; McDonald's omega; inverse Wishart distribution

Reliability analysis aims to disentangle the amount of variance of a test score that is due to systematic influences (i.e., true-score variance) from the variance that is due to random influences (i.e., error-score variance; Lord & Novick, 1968). The most straightforward way to quantify the proportion of true-score variance is by correlating two administrations of the same test to the same group of people under exactly the same conditions. By definition, this correlation equals the test-score reliability. Thus, reliability provides an idea as to what happens if a test is readministered. That is, what results would be obtained if one could replicate the measurement procedure in the same group as if nothing had changed compared to the first measurement, except for changes in random error.

Reliability is a general concept that is relevant across a range of different designs, such as test-retest, inter-rater, and single-test designs. In practice, researchers often have data from only a single test administration rather than multiple administrations, and thus have to estimate reliability from the data collected by means of a single test version.

A reliability analysis serves three purposes. First, researchers wish to confirm that their measures are reliable. To do so, practitioners usually choose an estimator, commonly coefficient α (Cronbach, 1951) and obtain a point estimate, which is compared to a cutoff value to determine bad, sufficient, or good reliability (Oosterwijk et al., 2019). Second, reliability, by itself, does not inform us about the precision with which an individual is measured. For that, one needs the standard error of the unobservable individual's distribution of test-score replications. In practice, one uses the standard error of measurement (*sem*) for this purpose. The *sem* indicates the standard deviation of the measurement error in the group of interest and is often interpreted as an estimate of measurement precision (Mellenbergh, 1996). The *sem* is calculated as $sem = SD(X)\ \sqrt{1-\rho}$, where $SD(X)$ is the standard deviation of the test score X and ρ is the reliability of the test score. Third, when assessing the correlation between two variables that are measured with error, researchers may estimate the correlation between the true scores by correcting for the unreliability of the test scores. This so-called correction for attenuation (Spearman, 1904) is common in validity research and is often used in meta-analyses (Schmidt & Hunter, 1999).

When one estimates a parameter, the point estimate can be accompanied by an uncertainty interval. Both a Bayesian credible interval (or credible interval

for short) and a frequentist confidence interval (or confidence interval for short) are representatives of an uncertainty interval, but stem from competing statistical frameworks (see Appendix A for a comprehensive description of the interval types used in this study). Unfortunately, substantive researchers almost always ignore these intervals and present only point estimates in the context of reliability analysis (e.g., Oosterwijk et al., 2019). This common practice disregards sampling error and the associated estimation uncertainty and should be seen as highly problematic. In this study, we show how the Bayesian credible interval can provide researchers with a flexible and straightforward method to quantify the uncertainty of point estimates in a reliability analysis.

We wish to demonstrate the benefits of conducting a reliability analysis in the Bayesian statistical framework. We show that the Bayesian reliability coefficients (a) perform equally well as their frequentist counterparts, and (b) provide an intuitive interpretation of uncertainty, allowing researchers to address questions that are beyond the scope of a frequentist analysis. We employ Bayesian estimation procedures for some of the most popular single-test reliability coefficients: coefficient α (Cronbach, 1951), coefficient λ_2 (Guttman, 1945), the greatest lower bound (glb) (Woodhouse & Jackson, 1977), and coefficient ω (McDonald, 2013). Coefficient α provides a lower bound to the reliability (Cronbach, 1951) and is by far the most frequently used reliability coefficient (Barry et al., 2014; Flake et al., 2017; Hogan et al., 2000). Coefficient λ_2 is another lower bound and is at least as high as α (Guttman, 1945); the glb is at least as high as coefficient λ_2 and comes closest to reliability among the lower bounds (Sijtsma, 2009). Finally, we consider coefficient ω as the most prominent representative of the factor analytic approach to reliability (Revelle & Zinbarg, 2009). To obtain a parsimonious and clear study design, we leave out other reliability estimators such as Revelle's beta (Revelle, 1979) or Guttman's other λ-coefficients (Guttman, 1945).[1] For a Bayesian solution of coefficients α, λ_2, and the glb we use and extend the methodology Padilla and Zhang (2011) described for the Bayesian estimation of coefficient α; we develop the first Bayesian version of coefficient ω. To promote their use, we implement the coefficients in an easy-to-use R-package.[2]

[1]Inclusion of coefficient λ_4, the maximum split half reliability, was considered but dismissed due to a strong positive bias and its uneconomic computational effort when the number of items exceeds ten.

[2]The R-package `Bayesrel` can be installed from CRAN or the latest version can be downloaded from https://github.com/juliuspf/Bayesrel.

The outline of this paper is as follows. First, we discuss the purpose of Bayesian reliability estimation and how the present approach relates to previous work. Second, we provide an overview of the current state of uncertainty estimation in reliability analysis. Third, we discuss reliability estimation in general and more specifically with respect to the different measurement models relevant to the different reliability coefficients. This results in the implementation of the Bayesian single-test reliability coefficients. Fourth, we apply the Bayesian coefficients both in a simulation study and to an example data set. Finally, we discuss our findings and their implications for future work on the topic of Bayesian reliability estimation.

Purpose of this work

Bayesian theory

In Bayesian inference, estimation uncertainty is quantified by a posterior distribution that represents the relative plausibility of different parameter values after the data have been observed. In order to obtain the posterior distribution we must first choose a prior distribution. The prior distribution represents the relative plausibility of different parameter values before the data have been observed. The prior distribution is then updated by means of the likelihood to yield the posterior distribution. Based on the posterior distribution we can report credible intervals that indicate the precision with which the parameter has been estimated, we can make statements about the probability that the parameter has a value larger or smaller than some threshold, and we can obtain a starting point for the analysis of additional data.

Previous work

Previous efforts to develop Bayesian reliability estimates focused on coefficient α. Li and Woodruff (2002) detailed a procedure to obtain the posterior distribution for α directly by using the distribution of the maximum likelihood estimator of the coefficient (Van Zyl et al., 2000). A similar approach was taken by Najafabadi and Najafabadi (2016), who employed an exact distribution function of coefficient α (Kistner & Muller, 2004) to estimate the posterior distribution of the coefficient.

In contrast, Padilla and Zhang (2011) used the posterior distribution of the covariance matrix of multivariate observations to calculate a Bayesian coefficient α, including credible intervals. Padilla and Zhang noticed that coefficient α is fully determined by the

covariance matrix, and that the posterior distribution of the covariance matrix takes a convenient form when one assumes the data to be multivariate normal. They validated the method in a small simulation study by measuring the bias of the Bayesian point estimate and the coverage of the credible intervals.

Novel work

In this study we use and extend the approach from Padilla and Zhang (2011). Their procedure of sampling the posterior covariance matrix can be extended to estimate a series of well-known coefficients in addition to coefficient α. Specifically, we use Padilla and Zhangs' approach to obtain Bayesian estimates of coefficient λ_2 and the glb.

Moreover, to the best of our knowledge, our development of a Bayesian coefficient ω is novel. The setup of ω is more complex than sampling from the multivariate normal, since one first needs to fit a factor model and then calculate ω from the model parameters. Fortunately, Lee (2007) detailed a sampling scheme to estimate a Bayesian single-factor model. Our work is the first to employ Lee's procedure and obtain a Bayesian estimate of coefficient ω, which is currently a popular reliability measure.

With this work we want to achieve two things. First, we wish to demonstrate the adequacy of Bayesian single-test reliability estimates by comparing them to their frequentist counterparts in a simulation study, and by explaining their benefits with an exemplary real-data set. Second, a long-standing problem in methodological research is the gap between theory and practice (Sharpe, 2013). We attempt to bridge this gap for Bayesian reliability analysis by introducing an R-package that contains the proposed methodology. This gives researchers all they need to conduct a Bayesian reliability analysis with multiple estimators and thus make more informed and intuitive inferences about reliability.

Current state of uncertainty estimation in reliability analysis

Researchers usually report a point estimate as an indicator for the quality of a research instrument but rarely report uncertainty intervals (Flake et al., 2017; Moshagen et al., 2019; Oosterwijk et al., 2019).[3] The almost complete absence of uncertainty interval reporting is surprising, especially in the light of multiple calls to improve the quality of psychological research by making increased use of confidence intervals (e.g., American Psychological Association, 2010; Association for Psychological Science, 2018; Cumming, 2014; Task Force Research on Reporting of Research Methods in AERA Publications, 2006; Thompson, 2002; Wilkinson, 1999).

The practice of reporting only reliability point estimates can hardly result from a lack of available methods. For example, bootstrapping is a mathematically simple re-sampling procedure that can be used to estimate a confidence interval when the distributional properties of an estimator are mostly unknown (e.g., DiCiccio & Efron, 1996; Efron, 1979). The so-called bootstrap confidence interval offers a well-known approach to the uncertainty estimation in reliability analysis. Several studies examined bootstrap confidence intervals for coefficient α and coefficient ω. They found that the intervals performed satisfactory under general conditions (Kelley & Pornprasertmanit, 2016; Padilla et al., 2012; Padilla & Divers, 2016; Raykov & Shrout, 2002). In addition to bootstrapping, analytic approaches to construct confidence intervals for the coefficients α and ω have been developed, and are accessible in standard software packages (see for α: Bonett & Wright, 2015; Feldt et al., 1987; Revelle, 2019; for ω: Kelley, 2018; Kelley & Cheng, 2012; Padilla & Divers, 2016; Raykov, 2002). Indeed, an analytic confidence interval for coefficient α is available in SPSS (v25), although disguised as a confidence interval for an intraclass correlation coefficient that equals coefficient α under certain conditions.

The common use of cutoff values might be one explanation why methods for estimating confidence intervals are not commonly employed in reliability analysis. Cutoff values such as .70 or .80 are frequently applied to determine "sufficient" or "good" reliability (Nunnally & Bernstein, 1994; Schmitt, 1996). Adding a confidence interval clashes with the idea of a simple cutoff value. For example, suppose the reliability point estimate of a single test is .73, the cutoff value for comparison .70, and the confidence interval around the estimate is [.67, .79]; by evaluating the point estimate the cutoff is exceeded and a clear conclusion is reached. By evaluating the confidence interval, however, one cannot say with "complete" certainty that the reliability is greater than .70.

Moreover, the interpretation of confidence intervals is subject to a common and troubling misconception. Many practitioners assume that a specific confidence interval contains the true parameter value with X% probability, or that one can be X% certain that the

[3]Personal communications with the authors in the first two references.

interval contains the true parameter value. However, a confidence interval is an interval generated by a procedure that in repeated sampling has at least an X% probability of containing the true value, for all possible values of that parameter (e.g., Morey et al., 2016; Neyman, 1937). Accordingly, it is incorrect to infer that the population value of a reliability parameter from a single test administration lies within the confidence interval limits with a probability of X%. Thus, confidence intervals do not answer practically relevant questions about the confidence one can have in a reliability estimate. Credible intervals offer a solution to this predicament. An X% credible interval encloses the parameter of interest with X% probability.

In sum, researchers almost never report interval estimates for their reliability coefficients. In addition, the questions they may have (e.g., "what is the probability that $\alpha > .80$?", "what is the probability that $\alpha \in [.70, .90]$?", "what is the relative support for different values of α provided by the data?") fall outside the purview of frequentist inference. To remedy these issues, we adopted and developed Bayesian estimation procedures for the most common single-test reliability coefficients.

Reliability estimation

According to classical test theory (CTT) a test score consists of a true score (systematic influences) and an error score (random influences), which are assumed to be uncorrelated (Spearman, 1904). In CTT, an individual's true score is defined as the expected value of a distribution of test scores for that person, known as the propensity distribution. The test scores are obtained in a hypothetical experiment, in which the same test is administered to the person under the same test administration conditions an infinite number of times. The crucial assumption is that the person remains the same from administration to administration. Consequently, the only source of variation in a person's test performance is random error (e.g., Sijtsma & Van der Ark, 2019).

Whereas the true score is defined for an individual, the true-score variance as well as the reliability are group characteristics. Reliability is traditionally defined as the product-moment-correlation between two parallel measures (Lord & Novick, 1968). Two test scores X and X' are parallel if: (1) the true scores T and T' are equal for the i-th individual, that is $T_i = T'_i$; and (2) $\sigma^2_X = \sigma^2_{X'}$, in the group of interest. Two test administrations are assumed parallel when the same test instrument is administered to the same sample of participants at two times while, hypothetically, the participants have no recollection of the previous administration. Then, the product-moment correlation between the two parallel test administrations equals the reliability, since both the true scores for individuals and thus the true-score variances as well as the test-score variances for the group are equal across administrations. It can be easily shown that for either of the parallel test administrations the proportion of test-score variance that is true-score variance is the same as the correlation of the test scores, which is the reliability (e.g., Lord & Novick, 1968):

$$\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X} = \frac{\sigma^2_{T'}}{\sigma^2_{X'}} = 1 - \frac{\sigma^2_E}{\sigma^2_X} = 1 - \frac{\sigma^2_{E'}}{\sigma^2_{X'}} \ , \qquad (1)$$

with σ^2_E denoting the error-score variance.

Since strictly parallel tests and parallel test administrations are unavailable in practice, and repeated test administrations often require substantial resources, a data matrix from a single test administration is commonly used to approximate the true-score variance (or complementary the error-score variance). Thus, reliability is calculated not as the product-moment correlation between scores across parallel test repetitions, but as a function of item covariances based on a single test administration. This makes it virtually impossible to accurately retrieve the true reliability, except when the items satisfy unrealistic conditions. Consequently, the following single-test reliability coefficients are only approximations to the true reliability.

CTT-coefficients

Coefficient α, coefficient λ_2, and the glb are based on CTT and are lower bounds to reliability (e.g., Guttman, 1945; Ten Berge & Zegers, 1978). To determine the error-score variance of a test, the coefficients estimate an upper bound for the error variances of the items. The estimators differ in the way they estimate this upper bound. The basis for the estimation is the covariance matrix $\boldsymbol{\Sigma}$ of multivariate observations. The CTT-coefficients extract information about error-score variance (or complementary, true-score variance) from $\hat{\boldsymbol{\Sigma}}$, which denotes the sample estimate of $\boldsymbol{\Sigma}$, and then weigh the estimated error-score variance against the total variance and subtract the result from 1 (or complementary, weigh the estimated true-score variance against the total variance). The CTT-coefficients estimate error-score variance from the variances of the items and true-score variance from the covariances of the items. The following paragraphs will elaborate on this.

Coefficient α

Let $\hat{\boldsymbol{\Sigma}}$ be the covariance matrix estimated from the data set consisting of k items, then

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\text{tr}(\hat{\boldsymbol{\Sigma}})}{S}\right), \tag{2}$$

where tr() is the sum of diagonal elements and S is obtained by summing all elements in $\hat{\boldsymbol{\Sigma}}$. Cronbach (1951) left his mark on coefficient α by discussing the measure in his famous article and deriving some well known and often repeated properties for it. Coefficient α equals reliability when the items are essentially tau-equivalent (Novick & Lewis, 1967).[4] When essential tau-equivalence does not hold, coefficient α is smaller than the true reliability, hence the lower bound property (Sijtsma, 2009; Zinbarg et al., 2005). Generally, coefficient α is closer to the reliability when a test is closer to unidimensionality (Dunn et al., 2014; Sijtsma, 2009).

Coefficient λ_2

Guttman proposed six lower bounds to reliability, one of which – λ_3 – equals coefficient α. These six coefficients take different approaches to the approximation of the error-score variance and are characterized by the quality that the reliability is never smaller than the largest of the six bounds (Guttman, 1945). The second lower bound is calculated as follows: We define $c = 1(\hat{\boldsymbol{\Sigma}} - diag(\hat{\boldsymbol{\Sigma}}))1'$ as the sum of squares of the off-diagonal elements in $\hat{\boldsymbol{\Sigma}}$ and S as before, then

$$\lambda_2 = \frac{S - \text{tr}(\hat{\boldsymbol{\Sigma}}) + \sqrt{\frac{k}{k-1}c}}{S}. \tag{3}$$

Coefficient λ_2 is always at least as large as coefficient α (Guttman, 1945; Sijtsma, 2009) and performed better than Guttman's other lower bounds (Oosterwijk et al., 2016).

Greatest lower bound

As its name implies, the glb is always at least as large as the other lower bounds. Following directly from the definition of the glb, the reliability lies in the interval [glb, 1]. The glb is calculated as follows: Let $\hat{\boldsymbol{\Sigma}} = \boldsymbol{C}_T + \boldsymbol{C}_E$ be the split of the sample covariance matrix into a matrix $\boldsymbol{C}_T$ that contains the true-score variances and a diagonal matrix $\boldsymbol{C}_E$ that contains the error-score variances. The proper estimates of $\boldsymbol{C}_T$ and $\boldsymbol{C}_E$ are found by maximizing the sum of the trace of $\boldsymbol{C}_E$ with the only condition being that $\boldsymbol{C}_T$ and $\boldsymbol{C}_E$ are positive semidefinite (e.g., Jackson & Agunwamba, 1977). Again, let S be the sum of all elements in $\hat{\boldsymbol{\Sigma}}$, then

$$\text{glb} = 1 - \frac{\text{tr}(\boldsymbol{C}_E)}{S}. \tag{4}$$

Finding the matrix $\boldsymbol{C}_E$ with maximum trace is not trivial. Various iterative matrix decompostion algorithms attempt to find a solution to the so-called "educational testing problem", all of which require a non-negligible amount of computational power (Bentler & Woodward, 1980; Ten Berge et al., 1981; Woodhouse & Jackson, 1977). Whereas among the lower bounds the glb comes closest to the true reliability, its sample estimate is prone to a capitalization on chance (Oosterwijk et al., 2016; Ten Berge & Sočan, 2004). This leads to a considerable positive bias of the estimate even up to sample sizes of $n = 1{,}000$ with the bias being smaller for fewer than ten items (Oosterwijk et al., 2016; Ten Berge & Sočan, 2004).

Factor model coefficient ω

Whereas coefficients α, λ_2, and the glb are rooted in CTT, coefficient ω is based on the single-factor model (McDonald, 2013). Specifically, the single-factor model assumes that one factor explains the covariances between the items (Spearman, 1904). The single-factor model corresponds to the congeneric measurement model, where all items load on one factor with varying loadings opposed to the tau-equivalence model where all loadings are assumed to be equal (e.g., Jöreskog, 1971).

The single-factor model can be considered a special case of a CTT-model, when one assumes that the common factor is a valid replacement for the true score. This way, the common factor variance replaces the true-score variance and the residual variances replace the error-score variance. Note that without this strong assumption coefficient ω could not be considered a reliability estimate. However, if the assumption holds, coefficient ω is a reliability measure for item sets. Please note that coefficient ω is not a measure of the fit of the single-factor model, it merely expresses reliability assuming unidimensionality.

Let

$$\boldsymbol{X}_{ij} = \lambda_j f_i + \boldsymbol{E}_{ij}, \tag{5}$$

where $\boldsymbol{X}_{ij}$ is the ith examinee's score on item j, λ_j are the loadings of the items on the common factor, f_i are

[4] The essential tau-equivalence condition states that all items of a test capture a participant's true score equally well (Lord & Novick, 1968, p. 50) – an assumption that can hardly hold in empirical research. Note that tau stands for the true score.

the examinees' factor scores, and $\boldsymbol{E}_{ij}$ is the ith examinee's error of item j that cannot be explained by the common factor. Further let $\boldsymbol{\Psi}$ be the diagonal variance matrix of $\boldsymbol{E}$ with ψ_j as the diagonal elements representing the residual variances of the items, then

$$\omega = \frac{(\sum \lambda_j)^2}{(\sum \lambda_j)^2 + \sum \psi_j}. \tag{6}$$

This equation coincides with Equation (6.20b) from McDonald (2013). Coefficient ω equals coefficient α (and the other lower bound coefficients) when the condition of tau-equivalence holds (Zinbarg et al., 2005). Coefficient ω represents a well studied alternative to coefficient α that shows good statistical properties (Kelley & Cheng, 2012; Zinbarg et al., 2005). To obtain the factor loadings and residual variances of the single-factor model one can apply a variety of application methods from the family of structural equation modeling (SEM). The most common are exploratory factor analysis (EFA), principal component analysis (PCA), principal factor analysis (PFA), and confirmatory factor analysis (CFA) (Revelle & Zinbarg, 2009; Zinbarg et al., 2006).

In the R-package, we implement both CFA and PFA to obtain coefficient ω. CFA models attempt to minimize the discrepancy between the model-implied covariance matrix and the sample covariance matrix using, for example, maximum likelihood or generalized least squares (Bollen, 1989). In a PFA, which is highly similar to the principal factor method or the principal axis method, the factor loadings are calculated from an altered sample covariance matrix by means of an eigendecomposition. The altered covariance matrix equals the sample covariance matrix with respect to the off-diagonal elements, but is different with respect to the diagonal, which contains the item communalities. The communalities are found in an iterative procedure that starts with the squared multiple correlations of the items (Rencher, 2002, p. 421 ff.).

Note that coefficient ω is sometimes split into ω_h (h for hierarchical) and ω_t (t for total) (Revelle & Zinbarg, 2009). In the calculation of these two coefficients a hierarchical multi-factor model replaces the single-factor model and additional factors account for group-specific variance. Subsequently, coefficient ω, as implemented in this work, can only be interpreted as a measure of reliability when the test is unidimensional and the single-factor model fully explains the true-score variance. The value of coefficient ω cannot address the question of dimensionality. To determine dimensionality one needs to apply more sophisticated factor analytic methods.

To summarize, the covariance matrix is sufficient for the CTT-coefficients, and the single-factor model parameters determine the value of coefficient ω. Thus, the obstacle in Bayesian single-test reliability analysis becomes the estimation of a posterior distribution for the covariance matrix and the estimation of the posterior distributions of the factor model parameters. Note that the formulas for the calculation of the coefficients remain unchanged in the Bayesian paradigm.

Bayesian single-test reliability estimation

Bayesian CTT-coefficients

We employ an approach to Bayesian reliability estimation that was described by Padilla and Zhang (2011), who provided a Bayesian version of coefficient α. Their approach is similar to the approach followed in this article and can be generalized to a series of different estimators.

Coefficients α, λ_2, and the glb are calculated on the basis of the sample covariance matrix. Thus, a straightforward way to obtain a posterior distribution of a CTT-coefficient is to estimate the posterior distribution of the covariance matrix and use it to calculate the estimate. In this procedure, we follow the methods also detailed in Murphy (2007). The author presents a simple and straightforward way to obtain a posterior distribution of the covariance matrix by choosing a conjugate prior distribution.

To facilitate the Bayesian estimation of reliability estimates, we assume data to be normally distributed with means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ for items $j = 1, \ldots, k$. Thus, the normal inverse Wishart distribution (NW^{-1}) is an obvious choice as a conjugate prior distribution: Let

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim NW^{-1}(\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Psi}, \nu_0) \ , \tag{7}$$

where $\boldsymbol{\mu}_0$ denotes the prior means, κ_0 is an inverse scaling parameter for the covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Psi} = (\frac{1}{\kappa_0}\boldsymbol{\Psi}_0)^{-1}$ is a positive definite inverse scaling matrix, and ν_0 are its degrees of freedom. The number of observations are n and the number of items are k. The prior hyperparameters are chosen as $\boldsymbol{\mu}_0 = \boldsymbol{0}, \kappa_0 = 10^{-10}, \boldsymbol{\Psi}_0 = \boldsymbol{I}_k$, and $\nu_0 = k$. This yields a relatively uninformative prior. The posterior distribution of the multivariate normal distribution is available in closed form:

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim NW^{-1}(\boldsymbol{\mu}_n, \kappa_n, \boldsymbol{\Psi}_n, \nu_n). \tag{8}$$

Since we are only interested in the covariance matrix, which is the basis for the CTT-coefficients, we sample the posterior covariance matrices from an inverse Wishart distribution, which equals:

$$\boldsymbol{\Sigma} \sim W^{-1}(\boldsymbol{\Psi}_n, \nu_n) \ , \tag{9}$$

with

$$\begin{aligned} \boldsymbol{\Psi}_n &= \boldsymbol{\Psi} + \boldsymbol{S} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^T \ , \\ \nu_n &= \nu_0 + n \ , \end{aligned}$$

where $\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ is the sum of squares matrix, $\boldsymbol{x}_i$ are the item scores of the ith examinee, and $\bar{\boldsymbol{x}}$ are the item means. The posterior distribution is obtained by drawing a sufficient number of random samples from the distribution denoted in Equation (9). By calculating any of the CTT-coefficients for each of the covariance matrices in the posterior sample, one obtains posterior samples for the reliability coefficients.

Bayesian factor model coefficient ω

Since coefficient ω is based on a factor model, its setup is more challenging. To obtain the posterior distributions of the parameters in the single-factor model, we need to sample from their conditional distributions. A comprehensive account to do so is given by Lee (2007), who describes how to obtain the posterior distributions of the quantities in Equation (5). We designed a Gibbs sampling algorithm to sample from the conditional conjugate distributions of the parameters in accordance with Lee (2007).

Recall the single-factor model from Equation (5) with ψ_j as the diagonal elements of $\boldsymbol{\Psi}$, the diagonal variance matrix of $\boldsymbol{E}$. We assume the following relatively uninformative conjugate prior distributions similar to Lee (2007, p. 71 ff.): Let $i = 1, \ldots, n$ observations and $j = 1, \ldots, k$ items, then $f_i \sim N(0, \phi)$; $\phi \sim W^{-1}(R_0, p_0)$ with scale matrix R_0 (in fact a scale value, as we only assume one factor) and degrees of freedom p_0; $\psi_j^{-1} \sim \Gamma(\alpha_0, \beta_0)$ with shape and rate parameter, also $\psi_j = 1/\psi_j^{-1}$, and $\lambda_j \sim N(0, \psi_j h_0)$. We fix the following prior hyperparameters: $R_0 = k, p_0 = k + 2$, $\lambda_{0_j} = 0, h_0 = 1$, $\alpha_0 = 2$, and $\beta_0 = 1$. Then the conditional posterior distributions are:

$$(f_i \mid X_{ij}, \psi_j, \lambda_j) \sim N \left[\left(\phi^{-1} + \lambda_j^T \psi_j^{-1} \lambda_j \right)^{-1} \lambda_j^T \psi_j^{-1} X_{ij}^T, \left(\phi^{-1} + \lambda_j^T \psi_j^{-1} \lambda_j \right)^{-1} \right] , \tag{10}$$

$$(\phi \mid X_{ij}, f_i) \sim W^{-1}(f_i^T f_i + R_0, n + p_0) \ , \tag{11}$$

$$(\psi_j^{-1} \mid X_{ij}, f_i) \sim \Gamma(n/2 + \alpha_0, \beta_j) \ , \tag{12}$$

$$(\lambda_j \mid X_{ij}, f_i, \psi_j) \sim N\left(m_j \sqrt{\phi}, \ \psi_j a \right) \ , \tag{13}$$

with

$$\begin{aligned} a &= \left(h_0^{-1} + f_i^T f_i \right)^{-1} \ , \\ m_j &= a\left(h_0^{-1} \lambda_{0_j} + f_i^T \mathbf{X}_j \right) \ , \\ \beta_j &= \text{diag}(\mathbf{B}) \ , \\ \mathbf{B} &= \beta_0 + \frac{1}{2}\left(X_{ij}^T X_{ij} - m_j^T a^{-1} m_j + \lambda_{0_j}^T h_0^{-1} \lambda_{0_j} \right) \ . \end{aligned}$$

In SEM, the latent factor needs to be assigned a metric, which is commonly done by fixing either the loading of one indicator to 1 or the factor variance to 1. Since we are interested in estimating the factor loadings we employ the latter procedure by dividing the factor scores drawn from Equation (10) by their standard deviation in every iteration. Subsequently, we multiply the means of the posterior loadings by the square root of the unstandardized factor variance (see Equation (13)) (Bollen, 1989).

Eventually, the iterative Gibbs-Sampling algorithm has the following steps:

1. Draw a sample value from one of the conditional distributions (henceforth the "first conditional distribution") in Equations (10)–(13) with starting parameter values drawn from the prior distributions.
2. Draw a sample value from any of the three other conditional distributions with the parameter values just drawn.
3. Draw a sample value from any of the two other conditional distributions with the parameter values just drawn.
4. Draw a sample value from the remaining conditional distribution with the parameter values just drawn.
5. Draw a sample value from the first conditional distribution with the parameter values just drawn.
6. Repeat steps **(2-5)** until sufficient samples have been drawn and the sampled values converged. The resulting posterior samples of λ_j and ψ_j can be used to calculate a posterior sample of coefficient ω by means of Equation (6).

Simulation study

We conducted a simulation study to evaluate the Bayesian single-test reliability coefficients across a variety of conditions and data sets. In this study, we compared the Bayesian coefficients with their frequentist counterparts and their population value.

Method

We constructed data generating covariance matrices for which we varied the average correlation between items ($\bar{\rho} = 0; .3; .7$) and the number of items ($k = 5; 20$). The covariance matrices were based on

the parameters of the single-factor model. Specifically, loadings and residual variances were combined to create the matrices by means of the equation $\boldsymbol{\Sigma} = \boldsymbol{\lambda}\Phi\boldsymbol{\lambda}' + \boldsymbol{\Psi}$, with loadings $\boldsymbol{\lambda}$, factor variance $\Phi = 1$, and diagonal residual variance matrix $\boldsymbol{\Psi}$. During the process the loadings and residual variances were sampled randomly until the correlation between items reached the desired average value. Consequently, the correlations between items varied in the medium and high-correlation conditions. In the high-correlation condition, the standard deviations of the average correlations ranged from .04 to .07; in the medium-correlation condition, they ranged from .17 to .21. In the zero-correlation condition the correlations varied slightly but were all very close to zero (ranging from .0002 to .0005). The resulting covariance matrices were used to generate multivariate normal data with means of zero in different sample sizes ($n = 50; 100; 500$).

We chose the zero-correlation condition to represent a state of very noisy data. The condition can be viewed as a baseline that is unlikely to appear in empirical research. However, by pushing the coefficients to their limits, we could examine how estimates behave for a wide variety of data sets. Also, in an extreme scenario like the zero-correlation condition, differences between the Bayesian and frequentist coefficients could emerge that we might otherwise not be able to detect.

The setup resulted in a total of 18 conditions with each condition being replicated 1,000 times. The population values of the CTT-coefficients were calculated from the data-generating covariance matrices. The population value of coefficient ω was computed from the factor loadings and residual variances that were used to construct the data-generating matrices. For the CTT-coefficients the data generation from a single-factor model can be considered proper, although one does not need to assume the data to be unidimensional to compute the coefficients.

For all analyses we used the R-package `Bayesrel` that we developed to facilitate the calculations and to make the Bayesian estimators accessible to other researchers. A detailed description of the package can be found in Appendix B. Although analytic solutions of confidence intervals for coefficients α and ω are available (e.g., Bonett & Wright, 2015; Kelley & Cheng, 2012), we used the non-parametric bootstrap (DiCiccio & Efron, 1996; Padilla et al., 2012; Padilla & Divers, 2016) to estimate percentile-type confidence intervals for reasons of consistency and comparability (see Table A1). To compute the confidence intervals, the number of bootstrap samples was set to 1,000. The number of iterations in the Bayesian solution was 1,000 with a burn-in of 50, because the sampling reached convergence very quickly. The credible intervals were highest posterior density (HPD) intervals (Box & Tiao, 1973) (see Table A1). When discussing the results from the simulation study, we refer to "percentile-type non-parametric bootstrap confidence intervals" as "confidence intervals" and we refer to "HPD credible intervals" as "credible intervals".

The frequentist calculation of coefficient ω was based on loadings and residual variances from a PFA, instead of a CFA, because – in several simulation runs – the CFA model-fitting did not converge or resulted in negative variances. Note that CFA and PFA yield virtually identical parameter estimates and thus nearly identical coefficient ω values when data are unidimensional.

We combined the posterior distributions of simulation runs for each condition by calculating the quantiles of those distributions and averaging each quantile. The analysis of the simulation results included a visual assessment of the consistency and the deviance of the estimates by means of summary plots, the coverage of the uncertainty intervals, the root mean-square error (RMSE) of the estimates, and the probability that a coefficient overestimated its associated population value (risk).

Coverage was computed as the percentage of simulation runs where the interval contained the population value. When calculating 95% intervals, we expected a well calibrated method to cover the population value in 95% of the cases. The RMSE quantifies the deviation of the posterior and bootstrapped samples of the estimators from their associated population values – averaged over simulation runs.[5] The RMSE can be interpreted on the same scale as the reliability and indicates how much an estimate spreads around its population value. Smaller RMSE values are associated with a less biased estimator. The "risk" was calculated by determining the quantile of the population coefficient value in the posterior distribution of the estimator. Let $F_X(x)$ be the cumulative distribution function of the posterior distribution of X evaluated at the point x, then r is the risk of overestimation: $r = 1 - F_X(x)$. F_X can be substituted by the posterior distribution of any one of the Bayesian reliability coefficients; x then becomes the associated population value of that coefficient. One would consider a coefficient to perform satisfactory if the risk of overestimation is close to .50, meaning the risk of underestimation is close to .50 as well. A more

[5]The RMSE is computed as $\sqrt{\sum_{t=1}^{T} \frac{(\hat{\theta}_t - \theta)}{T}}$, with T as the size of the posterior/bootstrap sample of a coefficient, $\hat{\theta}_t$ as the values of the posterior/bootstrap sample of a coefficient, and θ as the population value of a coefficient.

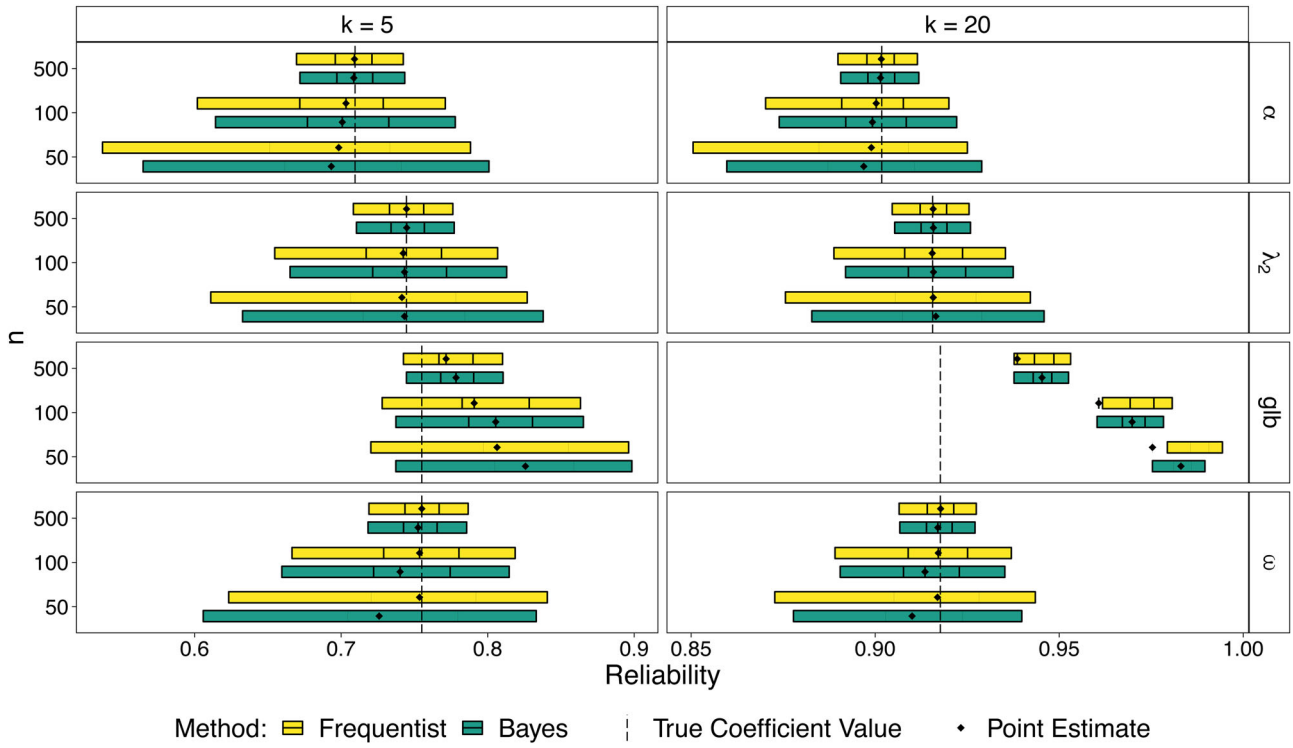

Figure 1. Simulation results for the medium-correlation condition. The endpoints of the bars are the 95% uncertainty interval limits. The 25%- and 75%-quartiles are indicated with vertical line segments.

conservative approach would also render values below .50 acceptable, since underestimating the reliability is more acceptable than overestimating it.

All results relate the estimated coefficients to each other or their corresponding population values. Thus, when we use the term bias, it reflects differences between the estimates and their population values.

Results

Comparison of frameworks

The results of the Bayesian coefficients were highly similar to the frequentist coefficients in almost all conditions. A summary of the results for the medium-correlation condition can be found in Figure 1. The figures summarizing the zero and high-correlation conditions can be found in Appendix C.

Figure 1 shows the results separated for the number of items (the columns) and the four coefficients (the rows). Each coefficient-block is divided into the different sample sizes (rows) and statistics frameworks (colors). The bars in the figure show the average 95% uncertainty interval limits, the vertical line segments in the bars indicate the average 25% and 75% quartiles, the dashed lines display the population values of the coefficients, and the average point estimates are denoted as rhombuses. We may notice, for example, that the average Bayesian point estimate for coefficient α with five items and 50 observations is slightly smaller than the frequentist one, and – for the same condition – the average confidence interval is shifted slightly more to the left and wider than the credible interval.

Except for λ_2 in the medium and high-correlation condition, the Bayesian point estimates over- or underestimated the associated population coefficient values slightly more than the frequentist ones when the sample size was small to medium in almost all conditions. This indicates an influence of the relatively uninformative prior distribution in small sample settings. The point-estimated Bayesian glb showed larger positive bias than the point-estimated frequentist glb across all conditions. In general, the Bayesian estimators converged to the associated population values with increasing sample size.

Both methodological approaches displayed similar interval coverage performance (see Table 1). The 95% intervals of both frameworks performed well in the medium and high-correlation conditions. The credible intervals for λ_2 performed slightly worse than the confidence intervals in the zero-correlation condition. The difference vanished with increasing associations between items. Furthermore, the credible intervals of coefficient ω performed satisfactory. In the zero-correlation condition the credible intervals for ω displayed better coverage than the confidence intervals with a small number of items.

Table 1. Coverage of the Population Values for the 95% Confidence and Credible Intervals.

			$\bar{\rho} \approx 0$		$\bar{\rho} \approx .3$		$\bar{\rho} \approx .7$		$\bar{\mu}$	
	k	*n*	Freq	Bayes	Freq	Bayes	Freq	Bayes	Freq	Bayes
		50	.927	.944	.932	.942	.936	.954		
	5	100	.938	.940	.936	.941	.943	.948	.939	.944
α		500	.942	.933	.945	.942	.950	.952		
		50	.905	.929	.927	.936	.925	.946		
	20	100	.934	.937	.934	.943	.943	.945	.934	.940
		500	.943	.943	.949	.939	.948	.946		
		50	.832	.784	.930	.928	.947	.951		
	5	100	.827	.789	.935	.937	.947	.948	.904	.893
λ_2		500	.823	.811	.943	.940	.954	.947		
		50	.839	.752	.941	.936	.938	.945		
	20	100	.843	.791	.940	.937	.946	.942	.909	.889
		500	.834	.809	.955	.947	.947	.945		
		50	.241	.349	.664	.557	.687	.570		
	5	100	.269	.418	.706	.646	.708	.625	.563	.560
glb		500	.313	.456	.742	.712	.739	.708		
		50	0	0	0	0	0	0		
	20	100	0	0	0	0	0	0	0	0
		500	0	0	0	0	0	0		
		50	.622	.977	.938	.940	.953	.941		
	5	100	.672	.986	.941	.947	.949	.954	.862	.956
ω		500	.781	.974	.945	.935	.953	.951		
		50	.998	.997	.941	.918	.938	.890		
	20	100	1	.995	.943	.931	.947	.926	.963	.950
		500	1	.997	.957	.949	.946	.944		

Note. The confidence intervals are percentile-type non-parametric bootstrap; the credible intervals are HPD intervals. A desirable coverage is close to .95.

In the zero-correlation condition, the RMSEs of the Bayesian ω were slightly smaller than the frequentist ω (see Tables C1 and C2). Other than that, we regard the differences in RMSEs between the Bayesian and frequentist coefficients as negligible.

Evaluation of coefficients

As expected – based on CTT-results – an increase in the number of items lead to better point estimation and narrower covering uncertainty intervals in the majority of the conditions. An exception was the glb, which displayed a large positive bias in both frameworks when the number of items was large – a finding consistent with previous research (e.g., Oosterwijk et al., 2017).

The coverage information underscored the poor performance of the glb. The uncertainty intervals of the glb – both confidence and credible intervals – did not cover the population value in any simulation run with 20 items. The bias of the estimator was so large that in the frequentist approach none of the 95% confidence interval limits enclosed the point estimate when the sample size was small or medium. This indicates that the bootstrap re-sampling of the data set lead to even more biased estimates for the glb. This happened to different degrees in all three correlation conditions and illustrates the capitalization on chance mechanism that affects the measure. In addition to the glb, coefficient λ_2 and frequentist ω displayed unsatisfactory coverage in the zero-correlation condition. The other estimators showed satisfactory coverage performance. Except for the glb, the other coefficients reached a coverage level close to 95% with increasing correlations. With the exception of coefficient α, all other estimators displayed worse coverage results in the zero-correlation condition compared to the other conditions.

The RMSEs of coefficient ω were smallest among the estimators in the zero-correlation condition. In the other correlation conditions coefficient α, λ_2 and ω reached similar RMSEs. Similar to the coverage results, the glb had the poorest RMSE values among the coefficients across all conditions.

We did not compute risk results for the zero-correlation condition because the population values of the coefficients were very close to zero (see Figure C1), and thus a reliability coefficient was in itself inclined to overestimate its true value, as negative values were deemed invalid. Thus, the risk results were only calculated for the medium and high-correlation conditions. Table 2 contains the exact values. Again, the glb was at high risk of overestimating the population coefficient value. Coefficients α, λ_2, and ω reached satisfactory results with ω showing a more conservative risk.

Table 2. Probability of overestimating the population coefficient value.

		$\bar{\rho} \approx .3$		$\bar{\rho} \approx .7$		
	n	k = 5	k = 20	k = 5	k = 20	$\bar{\mu}$
	50	.465	.457	.474	.437	
α	100	.468	.467	.485	.458	.474
	500	.497	.493	.492	.496	
	50	.533	.549	.513	.496	
λ_2	100	.517	.531	.510	.502	.517
	500	.513	.522	.506	.514	
	50	.867	1.00	.859	1.00	
glb	100	.852	1.00	.852	1.00	.925
	500	.835	1.00	.835	1.00	
	50	.428	.426	.374	.343	
ω	100	.443	.442	.410	.392	.429
	500	.479	.483	.462	.466	

Note. The probability is calculated by computing the quantile of the population coefficient value in the posterior distribution of its associated estimator. The values are averaged over simulation runs. A desirable result is a value close to .50 or below.

Example: eight-item questionnaire data

To illustrate the advantages of the Bayesian single-test reliability coefficients, we used a data set by Cavalini (1992) measuring coping style in response to malodorous environments. Both Sijtsma (2009) and Revelle and Zinbarg (2009) used the data to discuss several reliability coefficients. The data set consists of eight-item questionnaire data filled out by 828 participants. The questionnaire is Likert-scaled with scores from 0 to 3, but is usually treated as quasi-continuous. The data covariance matrix can be found in Appendix D. The data set is included in the R-package under the name "cavalini". A table containing the exact values of the point estimates and the uncertainty interval limits for both the frequentist and Bayesian framework together with the R-code to reproduce the results for the example data set can be found in Appendix D.

A simple research question for the reliability analysis of this data set might be: "Does the reliability of the test exceed a cutoff value of .80, hence is the reliability good?". In the current state of substantive research, the common way to answer this question would be to use the point estimate of coefficient α, $\hat{\alpha}_{freq} = .778$. Based on this result, we would conclude that the scale is associated with a sufficient but not a good reliability, depending on the specific cutoff criteria applied. Suppose we are in doubt about using the proper coefficient, because we are aware of some critiques about α and know that λ_2 is generally a better coefficient. Also, we heed recommendations about good research practices and decide to provide a confidence interval. Thus, in a more sophisticated analysis we calculate: $\hat{\lambda}_{2_{freq}} = .785, 95\%$ CI $[.758, .809]$. The correct interpretation of the 95% confidence interval is: If we would repeatedly draw samples from the same population and calculate the 95% confidence intervals for λ_2 in the same way each time, the true value would fall into the interval in 95% of the cases. Thus, we are unable to make any inference about the reliability lying in the particular confidence interval we just obtained.

In contrast, from the posterior distribution we can conclude that the specific credible interval contains 95% of the posterior mass. Since $\hat{\lambda}_{2_{Bayes}} = .784, 95\%$ HDI $[.761, .806]$, we are 95% certain that λ_2 lies between .761 and .806. Yet, how certain are we that the reliability is larger than .80? Using the posterior distribution of coefficient λ_2, we can calculate the probability that it exceeds the cutoff of .80: $p(\lambda_2 > .80 \mid \text{data}) = .075$. We can also determine the posterior probability of coefficient λ_2 being larger than .70, which is another common cutoff value, and smaller than .80, which, in this example, is $p(.70 < \lambda_2 < .80 \mid \text{data}) = p(\lambda_2 > .70 \mid \text{data}) - p(\lambda_2 > .80 \mid \text{data}) \approx 1 - .075 = .925$. In addition to making probabilistic statements about the values of the reliability coefficients, the Bayesian approach allows for prior knowledge to be incorporated into the analysis. For example, a data covariance matrix of a previous study, which worked with the same test instrument, can be used to update the prior distribution and enhance the precision of the subsequent reliability analysis.

Figure 2 displays the Bayesian prior and posterior distributions of the coefficients for the Cavalini-data set with the original sample size of 828 and a reduced sample size of 100 randomly drawn observations. The plots differ for sample sizes and show that the point estimates generally concur, but the posterior distributions and the width of intervals are quite different for the smaller sample size. While the Bayesian posterior distribution can be used to calculate probabilities of interest, its simple graphical display encourages the users' appreciation for the uncertainty of reliability estimates.

Finally, the R-package `Bayesrel` allows the display of the graphical posterior predictive check (PPC) for the single-factor model. The idea is to use the posterior predictive distributions of the model parameters to check model fit. In the PPC, which is similar to a scree plot, we compare the posterior model-implied covariance matrices with the covariance matrix of the data to see if the parameters we sampled under the single-factor model appear similar to the parameters observed in the data (Gelman et al., 2004, chapter 6.3). Since the model-implied covariance matrix is fully determined by the loadings and the residual variances, we can use the 95% limits of the loadings and residual variances to construct a lower and upper

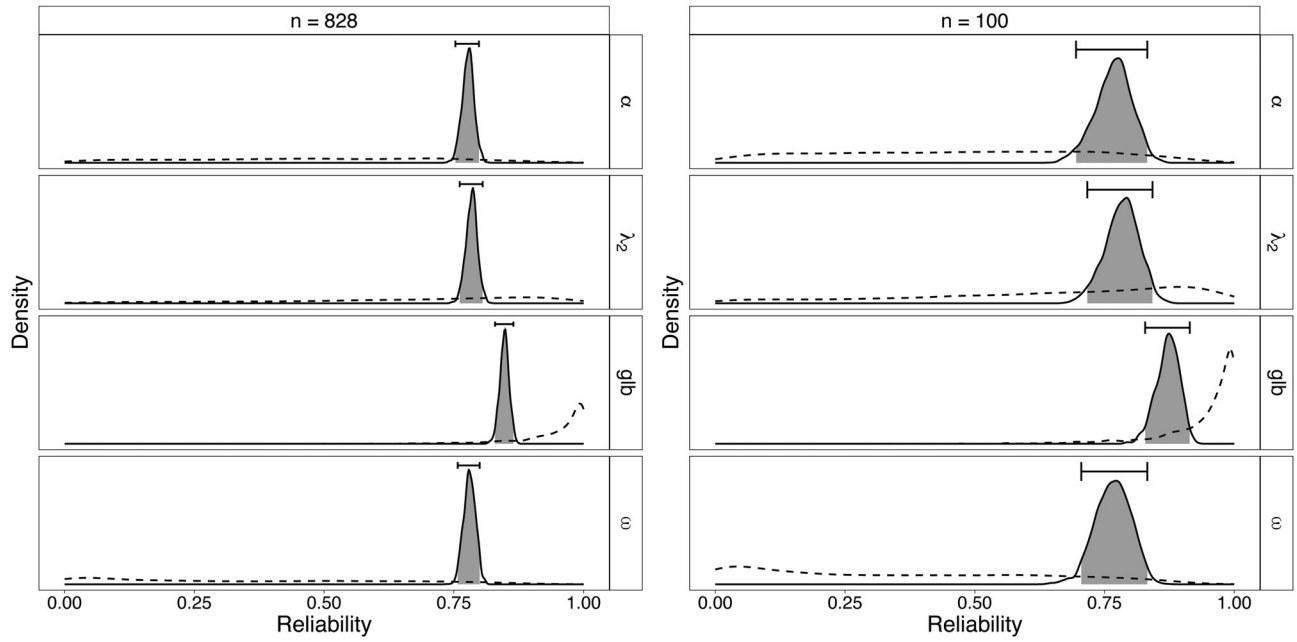

Figure 2. Bayesian results of empirical data set from Cavalini (1992) with eight items and sample size of $n = 828$, and $n = 100$ randomly chosen observations. Posterior distributions of estimators with dotted prior density and 95% credible interval bars.

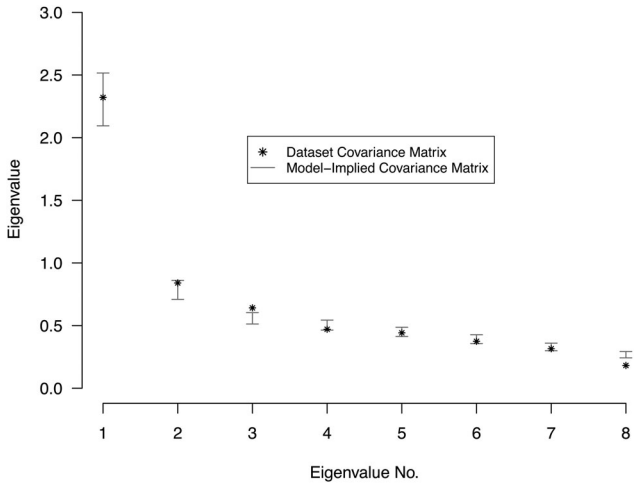

Figure 3. PPC of the single-factor model for the Cavalini-data from the Bayesrel-package. A good model fit implies that the gray bars enclose the black dots.

bound for the model-implied covariance matrix. We calculate the eigenvalues of the matrices and plot them (gray bars in Figure 3) together with the eigenvalues of the observed covariance matrix (black dots in Figure 3). If all gray bars enclose the black dots, we assume that the values sampled under the single-factor model are similar to the values of the observed data, hence the single-factor model fits the data. Inspecting the PPC for the full Cavalini-data set, we assess model fit to be mediocre at best, since not all of the observed eigenvalues are enclosed in the bars of the model predicted eigenvalues (see Figure 3).

In this instance, we are unable to interpret coefficient ω as a measure of reliability. This demonstrates that an analysis of dimensionality should always precede a reliability analysis and the PPC should only be used as a post-hoc check of model fit. If one had analyzed the dimensionality of the Cavalini data before conducting the reliability analysis, the results might have pointed toward a multidimensional scale. Thus, one could have divided the test into subscales and computed the reliability for each subscale separately. This way, the coefficients α, λ_2, and ω would have been better approximations of the reliability of the total scores based on each of the item subsets.

Discussion

When estimating reliability coefficients, researchers (a) predominantly report coefficient α; (b) almost always report only point estimates; and (c) rarely if ever report Bayesian inference. Here, we addressed these issues by implementing Bayesian inference procedures for some popular single-test reliability coefficients. Using the posterior distribution, researchers obtain a direct and intuitive appreciation for the uncertainty in their inference. Moreover, the posterior distribution allows them to obtain answers to questions that fall outside of the standard frequentist framework. Specifically, researchers can obtain the probability that a reliability coefficient falls in any particular interval of interest and update this probability continuously, as more data become available.

We tackled the issue of lacking uncertainty estimation and reporting in psychological research by offering an easy-to-use statistical method readily available for empirical researchers. The credible intervals implemented in this study offer a simple and intuitive solution to account for the sampling error in reliability estimation. We presented Bayesian procedures to estimate coefficient α, coefficient λ_2, the glb, and coefficient ω, and implemented them into an openly accessible R-package. A simulation study and the analysis of an exemplary data set demonstrated the adequacy and benefits of the Bayesian estimators. In addition, the study validated bootstrap confidence intervals for coefficient λ_2 and the glb.

In particular, the simulation study showed that the Bayesian coefficients converge to their population values and concur with the frequentist estimates. As expected, performance of the Bayesian coefficients improved with increasing sample size. Whereas the frequentist and Bayesian methods differed slightly with respect to point estimation, they generally coincided with respect to interval estimation in terms of coverage performance.

In contrast to the frequentist coefficients, the Bayesian reliability estimates allow for more intuitive and simple statements about reliability, which makes them superior in almost every aspect. The covariance sampling procedure introduced here can be easily extended to other estimators as long as they are calculated from the covariance matrix. Measures that are based on the repeated administration or the repeated rating of the same test, can be subject to future work on Bayesian test-retest or inter-rater reliability.

Except for λ_2, the performance of the Bayesian coefficients appears unsatisfactory for small samples. However, we do not consider this problematic. In Bayesian statistics the initial confidence in different values of a parameter is expressed by means of a prior distribution. The observed data then cause this confidence to be reallocated following Bayes' rule: Confidence is gained for parameter values that accord well with the data, and confidence is lost for values that accord poorly. The more informative the data, the larger the shift in confidence. Small samples generally result in modest changes of the prior distribution. Consequently, in practical applications with small sample data, a visual inspection of the posterior distribution would reveal that it is relatively wide, and the primary concern will focus on the large posterior uncertainty, not on how close the posterior mean may be to the true value. In addition, with small sample sizes the relatively uninformative priors for coefficient α, λ_2, and ω (e.g., see Figure 2) yield underestimates of the population values. In other words, the bias is conservative and researchers are safeguarded from overestimating reliability when the evidence in the data is small. When sample size is small, we urge researchers to try and collect more observations. With more incoming evidence the influence of the prior distribution will diminish and subsequently the Bayesian sample estimates will be closer to their respective population values.

More generally, the results for Bayesian parameter estimation are relatively robust to changes in the prior distribution, a regularity captured by the maxim "the data overwhelm the prior" (e.g., Wrinch & Jeffreys, 1919). Exceptions to this rule occur when the prior distribution is highly informative or when the sample size is very small. In the former case one may wonder why, with such firm knowledge already in hand, one would seek to collect additional data at all; in the latter case, it is prudent to interpret any strong conclusions with considerable caution. If researchers decide to employ informed prior distributions, these should be well motivated, for instance by an analysis of previous empirical findings or by a systematic elicitation effort (e.g., Stefan, Evans, & Wagenmakers, in press, and references therein). Even in these cases we see value in reporting the results for default priors as well, because these provide a reference against which to assess the conclusions from informed priors. Specifically, it is always possible to conduct a sensitivity analysis, where one examines the extent to which the conclusions vary as a result of reasonable changes in the prior distribution.

Some researchers may want to capitalize on the benefits of the Bayesian framework and incorporate informed prior knowledge into their analysis. This is not a trivial task and we caution readers against using more informative priors until more research into the topic is conducted. For academic purposes, however, we briefly outline a potential process to do so. One could use the posterior distribution of the covariance matrix from a previous similar data analysis as the prior distribution for the current analysis. For example, one wants to conduct a reliability analysis for a translation of the Cavalini-questionnaire administered in Germany. The same questionnaire in Dutch was administered in a previous study, and the authors of the study provided the sample covariance matrix. To obtain more precise CTT-estimators for the administration of the German questionnaire one could alter the parameters of the prior inverse Wishart distribution so that its mean equals the sample covariance matrix of the Dutch questionnaire. To obtain coefficient ω for

the German questionnaire one could adjust the hyperparameters of the prior distributions of the loadings and residual variances to better resemble the loadings and residual variances from a single-factor model fit of the sample covariance matrix of the Dutch questionnaire. With multiple similar questionnaires at hand, a more sophisticated approach is to apply a hierarchical model (e.g., Shiffrin et al., 2008) such that the estimate of reliability for a single small-sample questionnaire is informed by knowledge of the reliability for the other questionnaires, moderated by the extent to which the questionnaires are similar.

The Bayesian reliability coefficients

In view of common misconceptions about coefficient α (Cho, 2016; Cho & Kim, 2015; Hoekstra et al., 2019)[6]; we appeal to researchers to interpret α only as a lower bound to reliability. The Bayesian version of λ_2 performed equally well in the simulation study as coefficient α. We suggest that in the future researchers should take a closer look at coefficient λ_2, a greater lower bound to reliability than coefficient α (Guttman, 1945; Oosterwijk et al., 2016, 2017). Although the glb is the coefficient that comes closest to the true reliability on a population level (Ten Berge & Zegers, 1978), we recommend against using the glb in practice based on the performance in this study and others (Oosterwijk et al., 2016; Ten Berge & Sočan, 2004).

This study chose the same prior for the covariance matrices for all coefficients. Being sampled from an inverse Wishart distribution with the identity as a scaling matrix, the prior covariance matrices can be assumed relatively uninformative. This yielded acceptable flat priors for coefficients α and λ_2 with most values between 0 and 1 being approximately equally likely a priori. However, the prior of the glb was heavily skewed to the left. Specifically, the choice of a relatively uninformative prior for the covariance matrix lead to a prior distribution of the glb that put an unwanted amount of weight on values of .9 and above (see, e.g., Figure 2). This became more severe as test length increased. To establish the glb as a more popular measure of reliability, future work could investigate different prior distributions and their effect on the estimator.

The simulation results suggest that the Bayesian coefficient ω proved its usefulness as a reliability coefficient for unidimensional tests. Multidimensional tests, however, require a more sophisticated model, for example, a bi-factor model. For this, one needs to apply a more complex procedure of posterior sampling than the single-factor model (see e.g., Lee, 2007; Muthén & Asparouhov, 2012). Estimating coefficient ω_h (Zinbarg et al., 2005) from a Bayesian perspective can be a promising approach to indicate the measurement of a common factor in a multidimensional scale.

To summarize, the Bayesian reliability estimates displayed similar statistical properties as the frequentist estimators. Furthermore, the shape of the prior distribution may influence the outcome of a Bayesian reliability analysis when the sample size is small.

Limitations of the present study

Simulation

Our simulation study covered a limited range of conditions. First, we assumed data to be continuous and normally distributed. A Bayesian solution for data that are multivariate non-normal or even categorical is offered by Gaussian copula graphical models (Mohammadi, Abegaz, Heuvel, & Wit, 2017). Kelley and Pornprasertmanit (2016) discussed the implications of categorical data on different types of ω. Second, we only simulated data that were unidimensional. An account of the behavior of frequentist reliability coefficients with multidimensional data is given by Cho (2016) and in the case of coefficient ω by Kelley and Pornprasertmanit (2016). Third, we did not evaluate the coefficients with respect to the population reliability but with respect to their population values. The coefficients are approximations to population reliability, thus their population values are unequal to population reliability under realistic conditions. If one assumes that the factor model can substitute CTT for reliability analysis and thus the factor score of the single-factor model represents the true score, the population value of coefficient ω equals the population reliability. Since the purpose of our study was to compare the reliability coefficients of one statistical framework to another we refrain from comparing coefficients to population reliability.

Bayesian estimation

The Bayesian estimation procedure starts with the choice of adequate priors for the parameters. Given the sparsity of research on Bayesian reliability estimation, we conservatively decided on relatively uninformative priors on the covariance matrices and relatively uninformative priors on the factor model parameters. For future research an alternate approach

[6]The two most common errors among researchers about coefficient α are the failure to interpret α as a lower bound and the false assumption of α as an indicator for unidimensional data (Hoekstra et al., 2019).

would be to specify priors such that the resulting prior distribution of the reliability coefficients is uninformative, instead of the prior covariance matrix or the factor model parameters.

Conclusion

The posterior distribution answers practically relevant questions about the confidence one can have in reliability estimation. As such, the Bayesian estimation adds an essential measure of uncertainty to simple point-estimated coefficients. Adequate credible intervals for single-test reliability estimates can be easily obtained applying the procedure described in this article, and as implemented in the R-package Bayesrel. Whereas the R-package addresses substantive researchers who have some experience in programming, we admit that it will probably not reach scientists whose software experiences are limited to graphical user interface programs such as SPSS. For this reason we are currently implementing the Bayesian reliability coefficients in the open-source statistical software JASP (JASP Team, 2020).

Whereas we cannot stress the importance of reporting uncertainty enough, the question of the appropriateness of certain reliability measures cannot be answered by the Bayesian approach. No single reliability estimate can be generally recommended over all others. Nonetheless, practitioners are faced with the decision which reliability estimates to compute and report. Based on a single test administration the procedure should involve an assessment of dimensionality. Ideally, practitioners report multiple reliability coefficients with an accompanying measure of uncertainty, that is based on the posterior distribution.

Article Information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported in part by an NWO Vici grant (016.Vici.170.083) and an Advanced ERC grant (743086 UNIFY) to EJW.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design ….

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). American Psychological Association.

Association for Psychological Science. (2018). *Submission Guidelines*. https://www.psychologicalscience.org/publications/psychological_science/ps-submissions.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, *41*(1), 12–18. https://doi.org/10.1177/1090198113483139

Bentler, P., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*(2), 249–267. https://doi.org/10.1007/BF02294079

Bodnar, T., & Okhrin, Y. (2008). Properties of the singular, inverse and generalized inverse partitioned wishart distributions. *Journal of Multivariate Analysis*, *99*(10), 2389–2405. https://doi.org/10.1016/j.jmva.2008.02.024

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc.

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, *36*(1), 3–15. https://doi.org/10.1002/job.1960

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley.

Cavalini, P. M. (1992). *It's an ill wind that brings no good: Studies on odour annoyance and the dispersion of odorant concentrations from industries. (Unpublished* [doctoral dissertation]. University of Groningen, The Netherlands.

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

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational*

Research Methods, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Corrada Bravo, H., Borchers, B. (2020). Rcsdp: R interface to the CSDP semidefinite programming library [Computer software manual]. https://CRAN.R-project.org/package=Rcsdp (R package version 0.1.57.1)

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 189–212.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology (London, England : 1953)*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*(1), 93–103. https://doi.org/10.1177/014662168701100107

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. https://doi.org/10.1177/1948550617693063

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman and Hall.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology*, *22*(4), 351–364. https://doi.org/10.1080/13645579.2018.1547523

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*(4), 523–531. https://doi.org/10.1177/00131640021970691

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*(4), 567–578. https://doi.org/10.1007/BF02295979

JASP Team. (2020). *JASP (Version 0.12.2)[Computer software]*. https://jasp-stats.org/

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. https://doi.org/10.1007/BF02291393

Kelley, K. (2018). MBESS: The MBESS R package [Computer software manual]. R package version 4.8.0.

Kelley, K., & Cheng, Y. (2012). Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments. *Methodology*, *8*(2), 39–50. https://doi.org/10.1027/1614-2241/a000036

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. https://doi.org/10.1037/a0040086

Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, *69*(3), 459–474. https://doi.org/10.1007/BF02295646

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. John Wiley & Sons Ltd.

Li, J. C., & Woodruff, D. J. (2002). *Bayesian Statistical Inference for Coefficient Alpha*. (Tech. Rep.). ACT Research Report Series.

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

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*(3), 293–299. https://doi.org/10.1037/1082-989X.1.3.293

Mohammadi, A., Abegaz, F., Heuvel, E. V. d., & Wit, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(3), 629–645.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Moshagen, M., Thielmann, I., Hilbig, B. E., & Zettler, I. (2019). Meta-analytic investigations of the HEXACO Personality Inventory(-Revised): Reliability generalization, self-observer agreement, intercorrelations, and relations to demographic variables. *Zeitschrift Für Psychologie*, *227*(3), 186–194. https://doi.org/10.1027/2151-2604/a000377

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution* (Tech. Rep.). University of British Columbia.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Najafabadi, P. A. T., & Najafabadi, M. O. (2016). On the Bayesian estimation for Cronbach's alpha. *Journal of Applied Statistics*, *43*(13), 2416–2441. https://doi.org/10.1080/02664763.2016.1163529

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A*, *236*(767), 333–380.

Novick, M., & Lewis, G. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika*, *32*(1), 1–13. https://doi.org/10.1007/BF02289400

Nunnally, C. J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2016). Numerical differences between Guttman's reliability coefficients and the glb. In L. A. van der Ark, D. Bolt, W.-C. Wang, J. Douglas, & M. Wiberg (Eds.), *Quantitative*

psychology research: The 80th Annual Meeting of the Psychometric Society 2015, Beijing, China (pp. 155–172). Springer.

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2017). Overestimation of reliability by Guttman's λ_4, λ_5, and λ_6 and the greatest lower bound. In L. A. van der Ark, S. Culpepper, J. A. Douglas, W.- C. Wang, & M. Wiberg (Eds.), *Quantitative psychology research: The 81th Annual Meeting of the Psychometric Society 2016, Asheville NC, USA* (pp. 159–172). Springer.

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2019). Using confidence intervals for assessing reliability of real tests. *Assessment*, *26*(7), 1207–1216. https://doi.org/10.1177/1073191117737375

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: Coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement*, *76*(3), 436–453. https://doi.org/10.1177/0013164415593776

Padilla, M. A., Divers, J., & Newton, M. (2012). Coefficient alpha bootstrap confidence interval under nonnormality. *Applied Psychological Measurement*, *36*(5), 331–348. https://doi.org/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. https://doi.org/10.22237/jmasm/1304223840

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 195–212. https://doi.org/10.1207/S15328007SEM0902_3

Rencher, A. C. (2002). *Methods of multivariate analysis*. John Wiley & Sons, Inc.

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57–74. https://doi.org/10.1207/s15327906mbr1401_4

Revelle, W. (2019). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. https://CRAN.R-project.org/package=psych (R package version 1.9.12)

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*(3), 183–198. https://doi.org/10.1016/S0160-2896(99)00024-0

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

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*(4), 572–582. https://doi.org/10.1037/a0034177

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods.).*Cognitive Science*, *32*(8), 1248–1284. https://doi.org/10.1080/03640210802414826

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

Sijtsma, K., & Van der Ark, L. A. (2019). Reliability. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Springer International Publishing.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Task Force Research on Reporting of Research Methods in AERA Publications. (2006). Standards for reporting on empirical social science research in AERA publications: American educational research association. *Educational Researcher*, *35*(6), 33–40. http://journals.sagepub.com/doi/10.3102/0013189X035006033

Ten Berge, J. M., Snijders, T. A., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika*, *46*(2), 201–213. https://doi.org/10.1007/BF02293900

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625. https://doi.org/10.1007/BF02289858

Ten Berge, J. M., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*(4), 575–579. https://doi.org/10.1007/BF02293815

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32. https://doi.org/10.3102/0013189X031003025

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. https://doi.org/10.1007/BF02296146

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/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), 579–591. https://doi.org/10.1007/BF02295980

Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, *38*, 715–731.

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

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*(2), 121–144. https://doi.org/10.1177/0146621605278814

Appendix A

Uncertainty intervals

Table A1. Uncertainty intervals and their interpretation

Frequentist confidence interval	Bayesian credible interval
Definition An X% confidence interval for a parameter θ is an interval with limits (L, U) generated by a procedure that in repeated sampling has at least an X% probability of containing the true value of θ, for all possible values of θ (Morey et al., 2016; Neyman, 1937). **Procedure** One constructs a confidence interval around a sample estimate $\hat{\theta}$ by adding an estimate of uncertainty to $\hat{\theta}$. The estimate of uncertainty is obtained by means of the sampling distribution of the parameter θ. This can be done in several ways. We distinguish between two cases. One where the sampling distribution is known (analytic confidence intervals) and one where it is unknown (bootstrap confidence intervals). **Analytic confidence interval** For example, assume a participant gets 9 out of 12 true/false questions correct. One calculates an analytic confidence interval for the parameter θ (the probability of answering any one question correctly) on the basis of the known sampling distribution of θ, i.e., the binomial distribution. Commonly, one approximates the binomial distribution of θ with a normal distribution on the basis of the central limit theorem. Hence, the 95% analytic confidence interval is: $[(9/12) \pm 1.96 * \sqrt{.75 * .25/12}] = [.505, .995]$. **Bootstrap confidence interval** For some models, the sampling distribution may not be available or applicable. In these situations, bootstrapping can be used to obtain an empirical sampling distribution of the parameter θ by re-sampling the data with replacement and computing θ for each data sample (Efron, 1979). The "bootstrap" confidence interval can be constructed from this empirical distribution in a number of ways. **Percentile-type confidence interval** To obtain a percentile-type interval we discard $100(\alpha/2)\%$ of the mass in the tails of the empirical sampling distribution. Thus, the percentile-type interval is given by the quantiles of the empirical sampling distribution: $[\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2}]$, with, i.e., θ^* being the bootstrapped parameter estimates. For the binomial example the 95% percentile-type bootstrap confidence interval for θ equals [.5, 1]. Note that the percentile-type interval can be constructed for any sampling distribution (also a posterior distribution) and is not exclusively relevant to bootstrapped sampling distributions.	**Definition** An X% credible interval for parameter θ is an interval with limits (L, U) that encloses 95% of the posterior mass. **Procedure** One constructs a credible interval for a parameter θ by summarizing the posterior distribution of θ by means of an interval; we distinguish two types. **Central credible interval** Consider again the example about answering true/false questions, where a participant gets 9 out of 12 questions correct. One can obtain the posterior distribution of θ (the probability of answering any one question correctly) in closed form if one uses a beta (1, 1) distribution as the prior for θ. Then the posterior distribution of θ is a beta (10, 4) distribution. One can compute a central credible interval of the beta (10, 4) distribution by using the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution. This results in a 95% central credible interval of [.462, .909]. **HPD interval** Another popular way to construct a credible interval is to use the highest posterior density (HPD) region. The HPD interval is the smallest possible interval that contains $100(1 - \alpha)\%$ of the posterior mass (Gelman et al., 2004, chapter 2.3). For example, to obtain the interval limits of an HPD interval for the posterior probability of the binomial rate θ one could sample a sufficient number of values (e.g. 10,000) from the posterior distribution, i.e., the beta (10, 4), and compute the HPD interval with standard software packages. The 95% HPD interval then equals [.486, .926].

Appendix B

R-Package: `Bayesrel`

The functions for computing the frequentist and Bayesian reliability estimates were combined into the R-package `Bayesrel`. The purpose of `Bayesrel` is to provide six different single-test reliability measures in both the frequentist and Bayesian way (download: https://github.com/juliuspf/Bayesrel).[7] The frequentist ω-computation is by default a CFA. The glb is calculated based on the "glb.algebraic"-function from the psych package (Revelle, 2019) which uses the "csdp"-function to solve the educational testing problem (Corrada Bravo & Borchers, 2020).

`Bayesrel` provides the user with credible and confidence intervals – in addition to point estimates. In the latest version, the Bayesian point estimates are the means of the posterior distributions. The confidence intervals are by default calculated as percentile-type with the non-parametric bootstrap method (Efron, 1979). The credible intervals are the highest posterior density (HPD) intervals of the posterior samples.

In addition, the package:

- allows the calculation of the probability that a coefficient is larger than any specified value;
- displays a graphical posterior predictive check of the comparison between eigenvalues of the model implied covariance matrix and the sample covariance matrix to check the fit of the single-factor model;
- provides the commonly used "if-item-dropped" statistics.

Note that for the if-item-dropped statistics the partitioning of the posterior covariance matrices can be done without the need to re-sample, since the covariance matrices from multivariate normal data are inverse Wishart distributed (Bodnar & Okhrin, 2008). However, we recommend against using the if-item-dropped statistics to dismiss items without thorough theoretical considerations.

Appendix C

Additional simulation results

The aggregated simulation results for the zero and high-correlation conditions are shown in Figure C1 and C2, respectively.

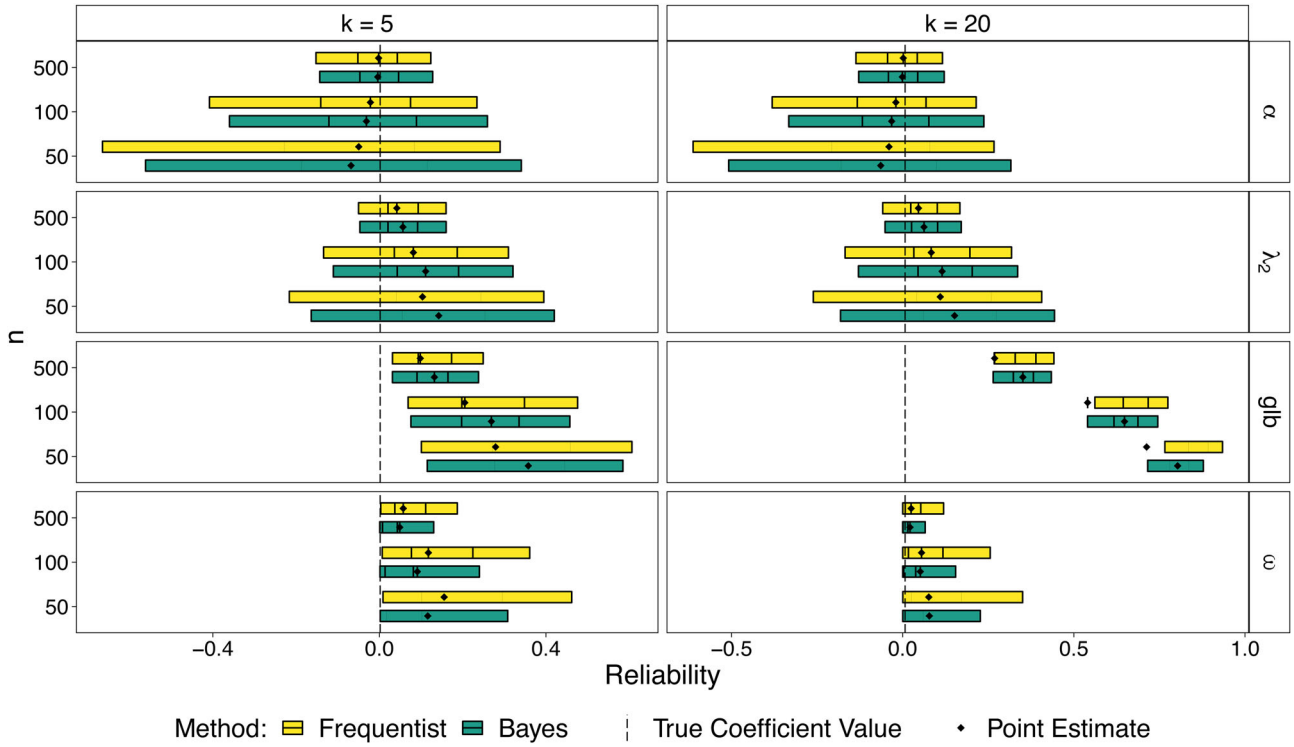

Figure C1. Simulation results for the zero-correlation condition. The endpoints of the bars are the 95% uncertainty interval limits. The 25%- and 75%-quartiles are indicated with vertical line segments.

[7] In addition to coefficient α, coefficient λ_2, the glb and coefficient ω, the package allows the calculation of coefficient λ_4 and λ_6. Users should beware of coefficient λ_4, however, since its calculation can take a considerable amount of time.

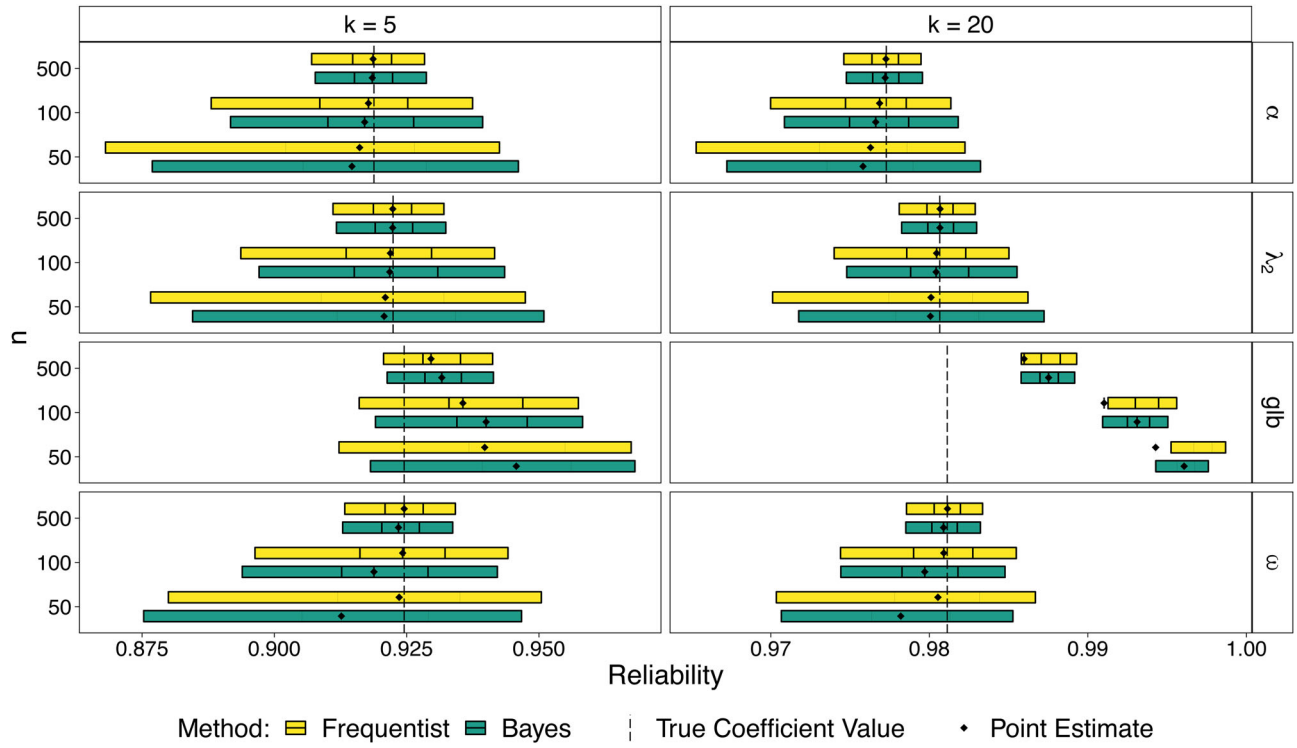

Figure C2. Simulation results for the high-correlation condition. The endpoints of the bars are the 95% uncertainty interval limits. The 25%- and 75%-quartiles are indicated with vertical line segments.

Below we calculated the root mean square errors (RMSE) of the values in the sampled distributions, posterior and bootstrap, and their associated population values – averaged over simulation runs. Thus, the RMSEs quantify the bias of the sampled coefficients in both frameworks. The tables are separated by number of items.

Table C1. RMSE of the posterior/bootstrap samples and the population coefficient value for 5 items.

		$\bar{\rho} \approx 0$		$\bar{\rho} \approx .3$		$\bar{\rho} \approx .7$		$\bar{\mu}$	
	n	Freq	Bayes	Freq	Bayes	Freq	Bayes	Freq	Bayes
	50	0.336	0.330	0.088	0.086	0.026	0.025		
α	100	0.225	0.223	0.059	0.059	0.017	0.017	0.098	0.097
	500	0.097	0.097	0.025	0.025	0.007	0.007		
	50	0.252	0.249	0.076	0.074	0.025	0.024		
λ^2	100	0.184	0.183	0.053	0.053	0.017	0.016	0.081	0.080
	500	0.089	0.088	0.024	0.024	0.007	0.007		
	50	0.390	0.379	0.091	0.088	0.028	0.027		
glb	100	0.295	0.289	0.066	0.065	0.020	0.020	0.120	0.117
	500	0.146	0.145	0.032	0.031	0.010	0.010		
	50	0.243	0.153	0.076	0.084	0.024	0.027		
ω	100	0.185	0.119	0.053	0.056	0.016	0.017	0.080	0.061
	500	0.092	0.064	0.024	0.024	0.007	0.007		

Note. Size of bootstrap and posterior samples is 1,000.

Table C2. RMSE of the posterior/bootstrap samples and the population coefficient value for 20 items.

		$\bar{\rho} \approx 0$		$\bar{\rho} \approx .3$		$\bar{\rho} \approx .7$		$\bar{\mu}$	
	n	Freq	Bayes	Freq	Bayes	Freq	Bayes	Freq	Bayes
	50	0.314	0.305	0.026	0.025	0.006	0.006		
α	100	0.210	0.206	0.017	0.017	0.004	0.004	0.075	0.073
	500	0.088	0.088	0.007	0.007	0.002	0.002		
	50	0.275	0.270	0.023	0.022	0.006	0.006		
λ_2	100	0.199	0.196	0.016	0.016	0.004	0.004	0.069	0.068
	500	0.093	0.093	0.007	0.007	0.002	0.002		
	50	0.855	0.794	0.070	0.065	0.016	0.015		
glb	100	0.674	0.641	0.055	0.052	0.013	0.012	0.230	0.218
	500	0.354	0.346	0.028	0.028	0.007	0.006		
	50	0.146	0.104	0.024	0.024	0.006	0.006		
ω	100	0.102	0.069	0.017	0.016	0.004	0.004	0.039	0.029
	500	0.044	0.027	0.007	0.007	0.002	0.002		

Note. Size of bootstrap and posterior samples is 1,000.

Appendix D

Cavalini data example

The covariance matrix of the Cavalini-data (Cavalini, 1992) is given by:

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{freq}} = \begin{pmatrix} 1.1101925 & .3847692 & .20299640 & .1693611 & .15245723 & .2230809 & .2640517 & .2634661 \\ .3847692 & .8941506 & .25851106 & .2030694 & .23972335 & .4532593 & .2337314 & .2447967 \\ .2029964 & .2585111 & .51025913 & .1219223 & .09646064 & .1766177 & .2038960 & .1396497 \\ .1693611 & .2030694 & .12192226 & .3954971 & .24639434 & .1906665 & .1407961 & .1266057 \\ .1524572 & .2397233 & .09646064 & .2463943 & .48975109 & .2925130 & .1450035 & .1244706 \\ .2230809 & .4532593 & .17661766 & .1906665 & .29251295 & .8582210 & .1856252 & .1923079 \\ .2640517 & .2337314 & .20389599 & .1407961 & .14500348 & .1856252 & .6216798 & .2867153 \\ .2634661 & .2447967 & .13964974 & .1266057 & .12447061 & .1923079 & .2867153 & .7058806 \end{pmatrix}$$

The results for the example are:

Table D1. Reliability statistics for cavalini-data.

		2.5% lower bound	point estimate	97.5% upper bound
α	Freq	.755	.778	.800
	Bayes	.753	.777	.798
λ_2	Freq	.758	.785	.809
	Bayes	.761	.784	.806
glb	Freq	.825	.845	.867
	Bayes	.829	.847	.865
ω	Freq	.760	.782	.805
	Bayes	.757	.780	.800

Note. The Bayesian point estimates are the means of the posterior distributions.

The R-code and output to reproduce the results for the example data:

```
> library(Bayesrel)
> set.seed(1234)
> res <- strel(data = cavalini)
> summary(res)
Call:
strel(data = cavalini)
Results:
```

```
              estimate   interval.low  interval.up
Bayes_alpha   0.777417   0.7529134     0.7984593
Bayes_lambda2 0.7842601  0.7611358     0.8055368
Bayes_glb     0.8473377  0.8292795     0.8648511
Bayes_omega   0.780281   0.757462      0.7997919
freq_alpha    0.7783201  0.7548938     0.7999913
freq_lambda2  0.7846576  0.7580393     0.8087328
freq_glb      0.8448238  0.825023      0.8667183
freq_omega    0.7820719  0.7595194     0.8046243
uncertainty interval: 0.95
```

```
> omega_fit(res)
   chisq       df    pvalue    rmsea rmsea.ci.lower
297.37364608 20.00000000 0.00000000 0.12942031
0.11663637
rmsea.ci.upper       srmr
0.14263586           0.06858548
```

See Figure 3.

```
   >#  give the probability for lambda2 larger than.80
> p_strel(res, "lambda2",.80)
     prior_prob      posterior_prob
     0.30366780      0.07481481
> # give the probability for lambda2 larger than.70
> p_strel(res, "lambda2",.70)
     prior_prob      posterior_prob
     0.4637053       1.0000000
>  # give the probability that lambda2 is larger
than.70 and smaller than.80
> p_strel(x, "lambda2",.70) - p_strel(x, "lambda2",.80)
     prior_prob      posterior_prob
     0.1600375       0.9251852
```