Scilight

Journal of Modern Applied Statistical Methods

https://jmasm.com

Article

Check for updates

Bayesian Estimation of Coefficient Alpha Using a Normal Posterior: Non-normal Distributions

John Mart V. DelosReyes * and Miguel A. Padilla

Department of Psychology, Old Dominion University, Norfolk, VA 23529, USA
* Correspondence: johnmart93@gmail.com

How To Cite: DelosReyes, J.M.V.; Padilla, M.A. Bayesian Estimation of Coefficient Alpha Using a Normal Posterior: Non-normal Distributions. *Journal of Modern Applied Statistical Methods* **2026**, *25*(1), 5. https://doi.org/10.53941/jmasm.2026.100005

Abstract: An alternative method for obtaining a Bayesian estimate for coefficient alpha by way of a posterior normal distribution was recently proposed. The performance of this method was initially assessed using Bayesian credible intervals (CrIs) via simulation under mundane conditions to establish baseline performance. It was found to generally have good performance under that set conditions. However, the performance of the alternative method using non-normal data was not assessed. Here, the alternative method is assessed using CrIs via simulation with a focus on non-normal data. Again, it was found that the alternative method generally had good performance. However, there were instances of poor performance when data came from simulation conditions that resulted in items being binary, non-normal, or not varying enough.

Keywords: coefficient alpha; Bayesian; non-normal; credible intervals

1. Introduction

Coefficient alpha is the most popular statistic used to estimate the reliability of items from a measurement instrument (e.g., test, survey, etc.) for a psychological construct. When considering a measurement instrument that contains k items, $x_1, x_2, \ldots, x_k$, coefficient alpha is defined as

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

where $\sigma_{x_{ii}}$ is the variance of item i and σ_X^2 is the total variance of the set of items for the measurement instrument [1]. Similarly, for a sample size n, the population parameters in Equation (1) can be replaced by sample estimates to obtain a coefficient alpha estimate as

$$\hat{\alpha}_c = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k}\hat{\sigma}_{x_{ii}}}{\hat{\sigma}_X^2}\right). \tag{2}$$

The enduring popularity of this statistic is rooted in three key qualities [2]. First, coefficient alpha is easy to calculate as it only requires the item covariance matrix to compute. In contrast, less popular forms of reliability are more difficult to compute and thus are not as often used. Second, coefficient alpha is viable for continuous, ordinal (i.e., Likert-type), and binary items. This allows coefficient alpha to be used with a variety of item response formats. Third, estimating coefficient alpha only requires one administration of a measurement instrument. This is a boon when most other forms of reliability require at least two administrations to estimate. The wide use of coefficient alpha in applied research has made coefficient alpha itself the focus of much research. Here, the focus will be on how non-normally distributed items impact the performance of credible intervals for a recently proposed Bayesian estimate of coefficient alpha.

1.1. Coefficient Alpha and Measurement Models

Coefficient alpha is an index of reliability but is only appropriate when items from a measurement instrument follow certain relationships as defined by measurement models from classical test theory (CTT; [3]). If the items have equal true scores and equal variability in measurement errors, items follow a parallel model. This results in items having a compound symmetric covariance matrix. If the items have equal true scores and unequal variability in measurement errors, items follow an (essentially) tau-equivalent model. This results in items having a covariance matrix that is compound symmetric with heterogeneous variances. If the items have unequal true scores and unequal variability in measurement errors, items follow a congeneric model. This results in items having a covariance matrix that is first-order heterogenous. For all measurement models, the measurement errors are assumed to be independent from the true scores and from the measurement errors of other items or separate occasions of the same item. These measurement models are important to consider as coefficient alpha is only equal to reliability (i.e., $\alpha_c = \rho_{xx'}$) when items follow a parallel or (essentially) tau-equivalent model. If the items follow a congeneric model, coefficient alpha can be shown to be only a lower bound to reliability (i.e., $\alpha_c \leq \rho_{xx'}$; [4]). As such, it is important to consider measurement model assumptions when using coefficient alpha as its interpretation changes depending on which measurement model the items follow.

1.2. Interval Estimation of Coefficient Alpha

Recent research on coefficient alpha has focused on the properties of its interval estimates. Interval estimation is of interest because it accounts for sampling error by way of estimating a range of possible values likely to contain a population parameter [5–7]. The most common type of interval estimate has been the confidence interval (CI). However, developing CIs for coefficient alpha has been challenging.

Much of the challenge in developing CIs for coefficient alpha has been related to trying to overcome assumptions related to estimating coefficient alpha limiting its applicability. One such assumption to address was assuming that items follow a parallel measurement model (i.e., items having a compound symmetric covariance matrix). An initial CI for coefficient alpha was proposed by Feldt [8] by deriving the sampling distribution of coefficient alpha assuming items were parallel and were normally distributed. Barchard and Hakstian [9] investigated the performance of this CI via simulation and found that the CI did not perform well if items were not parallel. This was significant as requiring items to be parallel is an assumption that is not typically met in applied settings. As a result, this early coefficient alpha CI was rarely used in applied settings [10]. Later, van Zyl et al. [11] showed that coefficient alpha in Equation (2) is a maximum likelihood estimate (MLE) with a normal asymptotic distribution. This coefficient alpha MLE assumes items are normally distributed but does not require items to be parallel (i.e., does not require items to have a compound symmetric covariance matrix). Duhachek and Iaobucci [10] would build off this finding to propose a normal theory (NT) CI for coefficient alpha. The significance of all this was that it led to developing coefficient alpha CIs that were not limited to the parallel item assumption though still required items to be normally distributed.

However, assuming items follow a normal distribution was also found to limit coefficient alpha CI applicability. Measurement instrument items used in the behavioral/social sciences tend to be Likert-type (or ordinal) or dichotomous (e.g., yes/no, true/false, etc.), making the normality assumption of items difficult to meet in practice. To help overcome this limitation, Yuan et al. [12] proposed a bootstrap CI for coefficient alpha. The advantage of the bootstrap method is that it makes no distributional assumptions as it uses resamples of the data (i.e., bootstrap samples) to generate an empirical sampling distribution (ESD) of a statistic from which inferences can be made. Additionally, the authors proposed a new, asymptotic distribution-free (ADF) CI for coefficient alpha, and compared the performance of the bootstrap, ADF, and NT CIs estimated from the Hopkins Symptom Checklist (HSCL; [13]). This comparison showed that the bootstrap CIs were most accurate. However, the ADF CI performed similarly (within three decimals) and were less computationally intensive than the bootstrap CIs, resulting in further investigation for the ADF CI [14]. Maydeu-Olivares et al. [14] compared the ADF CI against the NT CI via simulation and found that the ADF CI had better coverage probability (i.e., the proportion of estimated CIs that contained the true population coefficient alpha) than the NT CI when items were non-normal.

Further refinements in coefficient alpha CI research would take place. This refinement generally focused on the conditions used to determine the best performing CI. For example, Romano et al. [15] compared eight coefficient alpha CIs via simulation with a focus on dichotomous items. Here, it was found that the Fisher z-transformation and Bonett z-transformation CIs had the best coverage probability. Likewise, Padilla et al. [16] compared three bootstrap coefficient alpha CIs and four non-bootstrap coefficient alpha CIs via simulation with a focus on distribution types and item correlation types (i.e., measurement model types). In contrast to previous findings, the Fisher z-transformation, Bonett z-transformation, and NT CIs had the most instances of unacceptable

coverage probability. Additionally, they found that the NT bootstrap had the best coverage probability across the simulation conditions.

1.3. Performance of an Alternative Bayesian Estimation of Coefficient Alpha

The previously discussed literature outlined a brief history of CI development for coefficient alpha. These developments also contributed to a greater understanding of the coefficient alpha distribution and have been leveraged to approach coefficient alpha from a Bayesian perspective. This is significant as it allows coefficient alpha estimates to be updated with prior information as well as providing a more intuitive interpretation of interval estimates by way of Bayesian credible intervals (CrIs). Priors are of particular benefit in that they can be used to supplement parameter estimates when there is a small sample size [17]. Additionally, rather than interpret a 95% CI as 95% of all similarly constructed CIs containing the parameter of interest; the 95% CrI is instead interpreted as the parameter of interest as having a 95% chance of being within the bounds of the CrI given the data.

Recently, an alternative method to obtain a Bayesian estimate of coefficient alpha was proposed DelosReyes and Padilla [18]. This was done by leveraging the results from van Zyl et al. [11] (2000) wherein the MLE of coefficient alpha has a normal asymptotic distribution. From here, the posterior distribution for coefficient alpha was defined first by using conjugate priors such that the priors selected would result in a closed-form expression for the posterior. In cases without prior information, improper priors that do not integrate to 1 yet still result in posterior distributions that do integrate to 1 were selected. Given that this method is built on the van Zyl method, it is subject to its perks and penalties. For example, it does not require items to be parallel but assumes that items are normally distributed.

The authors assessed this alternative Bayesian estimate of coefficient alpha using three types of 95% CrIs via simulation [18]. The first was percentile-based, the second was NT-based, and the third was highest probability density (HPD) based. The simulation conditions included (a) number of items ($k = 5,$ 10, 15, 20), (b) number of item response categories ($M = 2,$ 3, 5, 7), (c) correlation type, and (d) sample size ($n = 50,$ 100, 150, 200, 250, 300). For correlation type, five different correlation structures were investigated. The first two correlation structures were from a parallel-item one-factor model with common loadings of $\lambda_1 = 0.55$ and $\lambda_1 = 0.705$. The last three correlation structures were generated from a congeneric-item one-factor model with loadings: $\lambda_3 = 0.2,$ 0.3, 0.4, 0.5, 0.6; $\lambda_4 = 0.3,$ 0.4, 0.5, 0.6, 0.7; and $\lambda_5 = 0.4,$ 0.5, 0.6, 0.7, 0.8. The investigated correlation structures resulted in population coefficient alphas ranging from 0.36–0.96. Items were binary or Likert-type and set to be (approximately) normal. For each simulation condition, 1000 replications were obtained. Priors used for the Bayesian estimate of coefficient alpha were set to be non-informative.

The CrIs investigated generally had acceptable coverage probability across the simulation conditions but tended to have instances of unacceptable performance when a sample size of 50 was paired with certain conditions. Notably, all CrIs had unacceptable coverage when there was a congeneric model with factor loading of $0.4 - 0.8$ paired with a sample size 50. In addition, the percentile and HPD CrIs had unacceptable coverage when there were 20 items paired with a sample size 50. Finally, the percentile CrI had unacceptable coverage when two item response categories were paired with a sample size 50. However, for all CrI methods, performance tended to be acceptable and stabilize to the target criteria of 0.95 as sample size increased.

The proposed Bayesian estimate of coefficient alpha had two noticeable features [18]. First, it was viable if sample size was large enough (i.e., $n > 50$). Second, it worked when items were parallel or congeneric. However, something to note was that this method was built upon the findings by van Zyl et al. [11] that assumed normally distributed items. Furthermore, the simulation used to assess the alternative Bayesian estimation method only used items that were (approximately) normal. Given that measurement instruments used in the behavioral/social sciences commonly use Likert-type (or ordinal) or dichotomous items, it is of interest to determine if this method based on (approximately) normal items works in explicitly non-normal scenarios and to what extent.

1.4. Bayesian Credible Interval Estimation

The CrI estimates for the present study are based on the Bayesian estimate of coefficient alpha as proposed by DelosReyes and Padilla [18]. The foundation of this method is built on the result from van Zyl et al. [11] who showed that

$$\hat{\alpha}_c \sim N\left(\alpha_c,\ \frac{\sigma_{\alpha}^2}{n}\right), \tag{3}$$

where

$$\sigma_{\alpha}^{2} = \left(\frac{k}{k-1}\right)^{2}\left[\frac{2}{(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1})^{3}}\right][(\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1})(tr(\boldsymbol{\Sigma}^{2}) + tr(\boldsymbol{\Sigma})^{2}) - 2tr(\boldsymbol{\Sigma})(\mathbf{1}'\boldsymbol{\Sigma}^{2}\mathbf{1})] \tag{4}$$

is the coefficient alpha variance, n is sample size, k is the number of items, $\boldsymbol{\Sigma}$ is a $k \times k$ item covariance matrix, and $\mathbf{1}$ is a conforming vector of ones. Here, the variance estimate (S_{α}^{2}) can be obtained by replacing $\boldsymbol{\Sigma}$ with the sample estimate $\boldsymbol{S}^{2}$ in Equation (4). By letting $\boldsymbol{y} = y_{1}, \ldots, y_{n}$ and using Bayes' rule for the normal distribution, it can be shown that

$$\pi(\alpha_{c},\ \sigma_{\alpha}^{2}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\alpha_{c},\ \sigma_{\alpha}^{2})p(\alpha_{c},\ \sigma_{\alpha}^{2})}{p(\boldsymbol{y})}. \tag{5}$$

To be specific, the posterior for coefficient alpha is described using conjugate priors. The conjugate prior for σ_{α}^{2} is

$$\sigma_{\alpha}^{2} \sim IG\left(\frac{\upsilon_{0}}{2},\ \frac{\upsilon_{0}\sigma_{0}^{2}}{2}\right), \tag{6}$$

where $IG(,)$ denotes an inverse-gamma distribution, and υ_{0} and σ_{0}^{2} can be thought of as the sample size and variance of a coefficient alpha from a prior set of observations, respectively. Following that, the conjugate prior for α_{c} is

$$\alpha_{c}|\ \sigma_{\alpha}^{2} \sim N\left(\alpha_{0},\ \frac{\sigma_{\alpha}^{2}}{\kappa_{0}}\right), \tag{7}$$

where α_{0} and κ_{0} can be thought of as the coefficient alpha and sample size from a prior set of observations, respectively. Taking these priors together, it can be shown that posterior for σ_{α}^{2} and α_{c} are as follows [18,19]:

$$\sigma_{\alpha}^{2}|\boldsymbol{y} \sim IG\left(\frac{\upsilon_{n}}{2},\ \frac{\upsilon_{n}\sigma_{n}^{2}}{2}\right), \tag{8}$$

where

$$\sigma_{n}^{2} = \frac{1}{\upsilon_{n}}\left[(n-1)S_{\alpha}^{2} + \upsilon_{0}\sigma_{0}^{2} + \frac{\kappa_{0}n}{\kappa_{n}}(\hat{\alpha}_{c} - \alpha_{0})^{2}\right], \tag{9}$$

$\upsilon_{n} = n + \upsilon_{0}$, $\kappa_{n} = n + \kappa_{0}$, S_{α}^{2} is the sample variance using Equation (4), and $\hat{\alpha}_{c}$ is the sample coefficient alpha. Furthermore, the posterior for α_{c} is

$$\alpha_{c}|\boldsymbol{y},\ \sigma_{\alpha}^{2} \sim N\left(\alpha_{n},\ \frac{\sigma_{\alpha}^{2}}{\kappa_{n}}\right), \tag{10}$$

where

$$\alpha_{n} = \frac{(\kappa_{0}/\sigma_{\alpha}^{2})\alpha_{0} + (n/\sigma_{\alpha}^{2})\hat{\alpha}_{c}}{\kappa_{0}/\sigma_{\alpha}^{2} + n/\sigma_{\alpha}^{2}} = \frac{n\hat{\alpha}_{c} + \kappa_{0}\alpha_{0}}{\kappa_{n}}. \tag{11}$$

If prior information is unavailable, improper priors (i.e., priors that do not integrate to 1 yet still result in posteriors that do integrate to 1) such as $p\left(\frac{1}{\sigma_{\alpha}^{2}}\right) \propto \sigma_{\alpha}^{2}$ and $p(\alpha_{c}) \propto 1$ can be used. This results in posteriors that are completely determined by the data such as

$$\sigma_{\alpha}^{2}|\boldsymbol{y} \sim IG\left(\frac{(n-1)}{2},\ \frac{(n-1)S_{\alpha}^{2}}{2}\right) \tag{12}$$

and

$$\alpha_{c}|\boldsymbol{y},\ \sigma_{\alpha}^{2} \sim N\left(\hat{\alpha}_{c},\ \frac{\sigma_{\alpha}^{2}}{n}\right). \tag{13}$$

For this study, three different CrIs are of interest to draw parity with the motivating study that introduced the Bayesian estimate of coefficient alpha detailed earlier. A percentile CrI can be obtained using the lower $\alpha/2$ and upper $1 - \alpha/2$ of $\pi(\alpha_{c}|y,\ \sigma_{\alpha}^{2})$. A NT CrI can be obtained with

$$\hat{\alpha}_{c} \pm z_{\alpha/2}SE(\hat{\alpha}_{c}), \tag{14}$$

where $SE(\hat{\alpha}_{c})$ is the standard deviation of $\pi(\alpha_{c}|\boldsymbol{y},\ \sigma_{\alpha}^{2})$. A HPD CrI can be obtained by letting a subset of the parameter space for $\hat{\alpha}_{c}$, $c \subset A$ be defined as

$$c = \{\hat{\alpha}_{c}: \pi(\alpha_{c}|\boldsymbol{y},\ \sigma_{\alpha}^{2}) \geq h\}, \tag{15}$$

where h is the largest number such that

$$\int_h \pi(\alpha_c|\boldsymbol{y},\ \sigma_\alpha^2)d\alpha_c = 1-\alpha. \quad (16)$$

In addition, an application example for these CrIs is included and discussed after the simulation results.

2. Methodology

2.1. Simulation

A Monte Carlo Simulation was used to investigate the properties of the Bayesian estimate of coefficient alpha and its corresponding CrIs. The simulation conditions were 4 (number of items) × 5 (correlation type) × 6 (sample size) × 4 (item response categories) × 3 (distribution type) for a total of 1440 conditions. All items were Likert-type (ordinal) or binary. For each simulation condition, 1000 replications were obtained. For consistency and comparison with other CrIs and CIs in the previous literature, non-informative priors as presented in DelosReyes and Padilla [18] were used. In addition, 2000 posterior draws were used.

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

1. Select the structure of the $k \times k$ correlation matrix **P**, where k is the number of items.
2. Select a set of thresholds $\boldsymbol{v}$ to categorize items to a predetermined skewness and kurtosis.
3. Generate an $n \times k$ multivariate data matrix $\mathbf{Z} \sim N(0, \mathbf{P})$, where n is the sample size.
4. Categorize the generated data **Z** using the thresholds in $\boldsymbol{v}$ to generate the dataset **X**. Each variable x in **X** is categorized by the thresholds as follows: $x = m$ if $v_m < z < v_{m+1}$ for $m = 0,\ 1,\ \ldots,\ M-1$, where $v_0 = -\infty$ and $v_M = \infty$, and M is the number of categories.
5. Compute the true population coefficient alpha (α_c) according to **P** and the thresholds in $\boldsymbol{v}$. Note that α_c is based on covariances after categorization. See Maydeu-Olivares et al. [14] for further details.
6. Estimate the Bayesian coefficient alpha CrIs from **X** as outlined in DelosReyes & Padilla [18].
7. Determine if the CrIs contain the true population coefficient alpha (α_c).

Details of the simulation conditions are given below.

2.2. Conditions

2.2.1. Number of Items (*k*)

Previous research on coefficient alpha has investigated the number of items ranging from 3 to 40 [9–11,14,16,20]. However, most of that previous research focused on items ranging from 5 to 20. Additionally, it is noted that going beyond 20 items for a measurement instrument reaches a point of diminishing returns with regards to coefficient alpha. Therefore, the following number of items were investigated: k = 5, 10, 15, 20.

2.2.2. Correlation Type (*P*)

Five different item correlation structures **P** were investigated. The first two correlation structures were from a parallel-item one factor model with common loadings $\lambda_1 = 0.55$ or $\lambda_2 = 0.705$. These two models will generate compound symmetric item correlation structures with $\rho = 0.30$ or 0.56, respectively. The next three correlation structures were generated from a congeneric item one-factor model with the following loadings: $\lambda_3 = 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6$; $\lambda_4 = 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7$; and $\lambda_5 = 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8$. These correlation structures were chosen as they were previously investigated and allow for a greater range of conditions to explore the impact on coefficient alpha CIs/CrIs [14,16]. Note that the loadings are standardized for these factor models. A summary for the population coefficient alpha these correlation structures produce is given in Table 1.

Table 1. Summary of resultant population coefficient alphas from correlation structure types.

Correlation Structure	Minimum α_c	Mean α_c	Maximum α_c
Parallel:			
$\lambda = 0.55$	0.48	0.75	0.89
$\lambda = 0.705$	0.71	0.88	0.96
Congeneric:			
$\lambda = 0.2,0.3,0.4,0.5,0.6$	0.30	0.59	0.78
$\lambda = 0.3,0.4,0.5,0.6,0.7$	0.42	0.71	0.86
$\lambda = 0.4,0.5,0.6,0.7,0.8$	0.55	0.79	0.91

Note. λ = standardized factor loadings.

2.2.3. Sample Size (n)

The following sample sizes were investigated: $n = 50,$ 100, 150, 200, 250, 300. This selection is in line with most recent research on coefficient alpha [10,14,16,20]. The sample size selections above 200 are beyond the point of diminishing returns for coefficient alpha but were investigated as they are typically found in behavioral/social science research.

2.2.4. Item Response Categories (M)

Previous research on coefficient alpha has investigated numbers of item response categories ranging from 2 to 7 [12,14–16]. To draw parity with previous research, the following common choices for response categories were investigated: $M = 2,$ 3, 5, 7. For each response category, the first category was set to 0. For example, for an item with seven response categories, $m = 0,$ 1, 2, 3, 4, 5, 6.

2.2.5. Distribution Type

Three different distribution types were investigated that are dependent on the number of item response categories (M). When $M = 2$ (i.e., items are binary), the thresholds for $\boldsymbol{\nu}$ were chosen such that the distributions had the following characteristics:

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

The type 1 and 2 distributions for binary items were studied by Padilla et. al. [16]. The type 3 distribution for binary items was studied by Maydeu-Olivares et al. [14]. When $M = 3,$ 5, 7 (i.e., $M > 2$) the thresholds for $\boldsymbol{\nu}$ were chosen such that the distribution had the following characteristics:

1. Type 1: skewness $= 0$ and kurtosis $= 0$
2. Type 2: skewness $= 0$ and kurtosis $= 0.88$
3. Type 3: skewness $= 0.97$ and kurtosis $= -0.20$

The type 1 distribution for $M > 2$ items was studied by Padilla et. al. [16]. The type 2 and 3 distributions for $M > 2$ items were studied by Maydeu-Olivares et al. [14]. Investigating this breadth of distribution types allows for the assessment of how the alternative Bayesian estimation of coefficient alpha responds to deviations from normality and draws parity with past research. The distribution types for items with two and five categories are presented in Figure 1. Additionally, Table 2 provides the thresholds used to produce these distributions.

Table 2. Thresholds used to categorize items to target distribution types.

Number of Items	Distribution Type		
	Type 1	**Type 2**	**Type 3**
Two	0	0.929	0.253
Three	−0.9675, 0.9675	−1.132, 1.132	0.24, 1.341
Five	−1.644, 0.642, 0.642, 1.644	−1.642, −0.842, 0.842, 1.642	0, 0.525, 1.038, 1.651
Seven	−1.81, −1.3, −0.3581, 0.3581, 1.3, 1.81	−2.2, −1.25, −0.832, 0.832, 1.25, 2.2	−2.25, −0.19, 0.43, 0.748, 1.106, 1.6

Note. For two items: Type 1 has skewness $= 0$, kurtosis $= -2$; Type 2 has skewness $= -1.70$, kurtosis $= 0.88$; and Type 3 has skewness $= 0.41$, kurtosis $= -1.83$. For three or more items: Type 1 has skewness $= 0$, kurtosis $= 0$; Type 2 has skewness $= 0$, kurtosis $= 0.88$; and Type 3 has skewness $= 0.97$, kurtosis $= -0.20$.

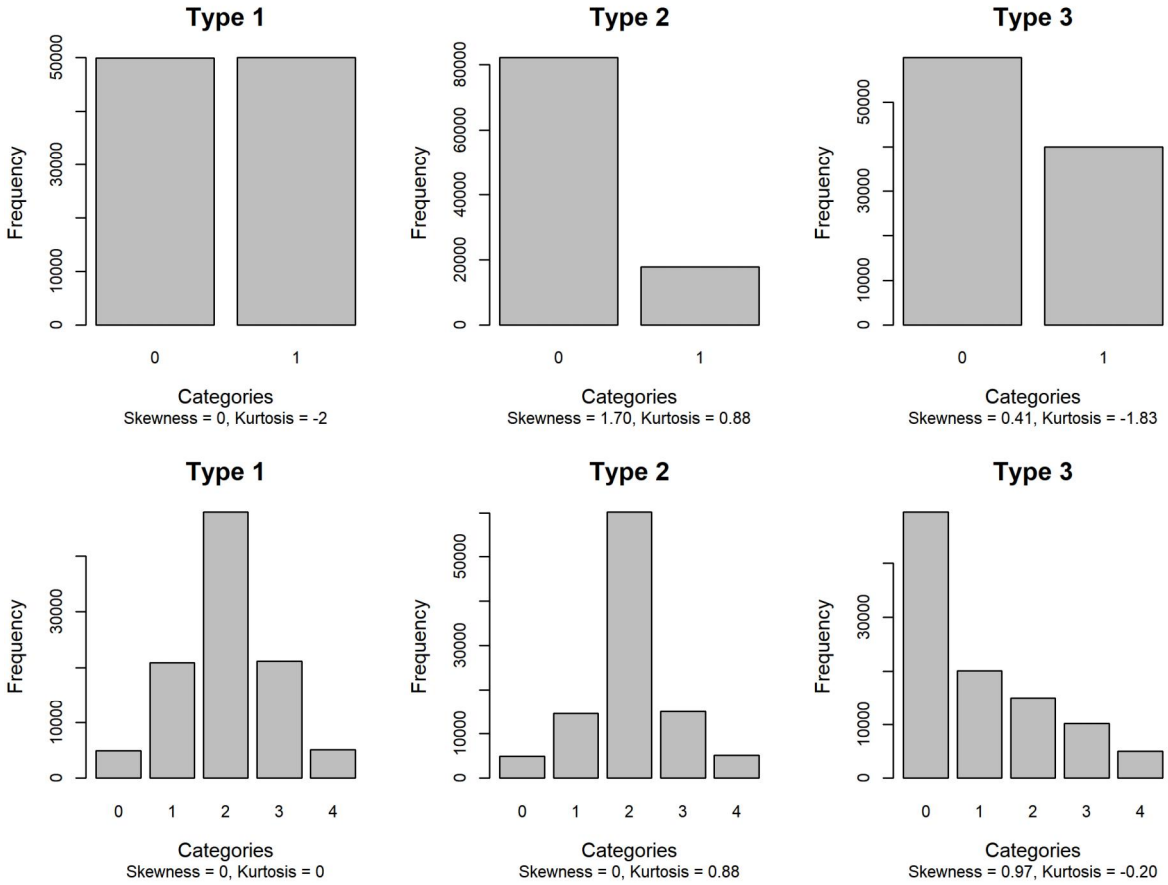

Figure 1. Bar graphs of the dichotomous and 5-point likert items utilized in simulation.

2.3. Analysis

To evaluate the performance of the CrIs, coverage probability was used. Coverage probability is defined as the proportion of estimated CrIs that contain the true population coefficient alpha. Coverage probability was assessed using Bradley's [21] liberal criterion defined by

$$1 - 1.5\alpha \leq 1 - \alpha^* \leq 1 - 0.5\alpha, \tag{17}$$

where α^* is the true type I error probability. Here the $100(1-\alpha)\%$ CrIs were estimated with $\alpha = 0.05$. Therefore, acceptable coverage probability is given by $[0.925,\ 0.975]$.

CrI width was also considered as a criterion for assessing CrI performance. CrI width gives information about the precision of a CrI estimate [22]. It is possible for two CrI methods to have the same coverage but with different interval widths (i.e., levels of precision). As such, CrI width is relevant if the CrI methods performed similarly in coverage probability.

3. Results

3.1. Coverage Probability

Figure 2 shows the coverage probability performance for each CrI across the simulation main effects (i.e., the simulation of conditions of *k*, *P*, *n*, *M*, and Distribution Type individually). It was found that all CrI methods performed similarly to one another and generally had acceptable coverage probability across all simulation conditions. However, there are some key patterns to notice across all CrI methods investigated. For the item response categories condition, there was more variability in coverage probability when there were only 2 categories compared to all other categories. In addition, for the distribution type condition, there was more variability in coverage probability in distribution type 2 compared to all other conditions. Lastly, for the sample size condition, coverage probability tended to be near the upper bound of the coverage probability criteria (i.e., 0.975) when the sample size was 50 and stabilized to 0.95 as sample size increased.

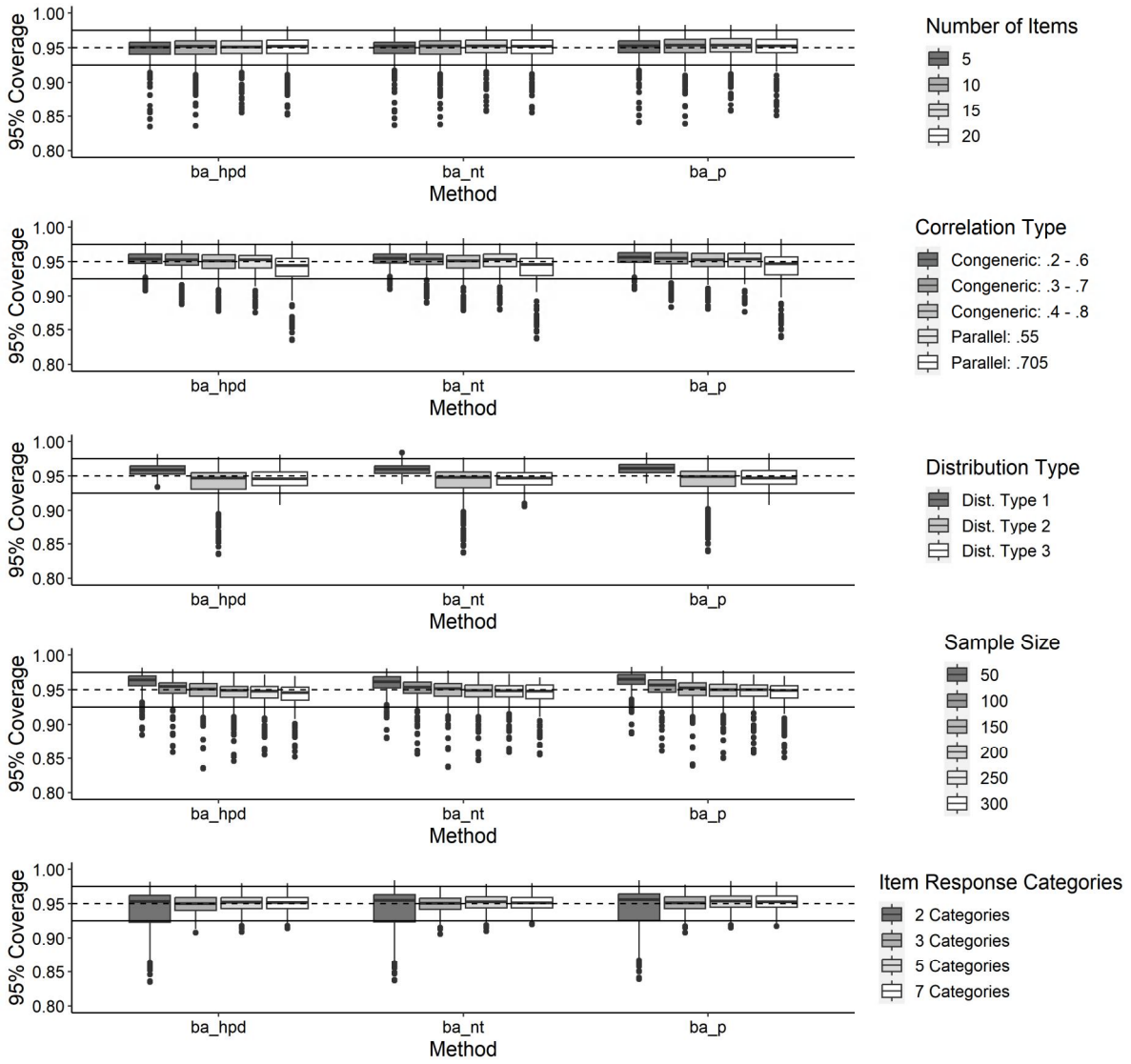

Figure 2. Distribution of the 95% credible interval coverage for all main effect simulation conditions. Note. Credible interval types denoted by the following: Bayesian alpha highest probability density (ba_hpd), Bayesian alpha normal theory (ba_nt), Bayesian alpha percentile (ba_p). Bayesian methods based on 2000 posterior draws. Dashed line at 0.95 and solid lines at acceptable coverage of [0.925, 0.975]. Correlation types indicate structure and range of loadings. Dist. = Distribution.

Figures 3–5 show the coverage probability performance of each CrI method based on the pairwise simulation conditions (i.e., any pair of combinations of the simulation conditions of *k*, *P*, *n*, *M*, and Distribution Type). Again, all CrIs performed similarly to one another and generally had acceptable coverage probability across all pairwise simulation conditions. However, there were some performance discrepancies amongst the CrI methods depending on the specific combination of pairwise simulation conditions.

For the percentile CrI (Figure 3), the simulation conditions that involved binary items ($M = 2$), distribution type 2, and/or a parallel model with common loadings of 0.705 tended to have unacceptable coverage. For binary items, performance was impacted across all number of items; for all correlation types except a congeneric model with factor loadings of $0.2 - 0.6$; or with a sample size of $150 - 300$. For distribution type 2, performance was impacted across all number of items; or with sample size of 100–300. For a parallel model with common loadings of 0.705, performance was impacted across all number of items; or with a sample size of $100 - 300$. Unsurprisingly, the pairwise combinations of binary items with distribution type 2; distribution type 2 with a parallel model with common loadings of 0.705; or binary items with a parallel model with common loadings of 0.705 had the largest performance impact of unacceptable coverage.

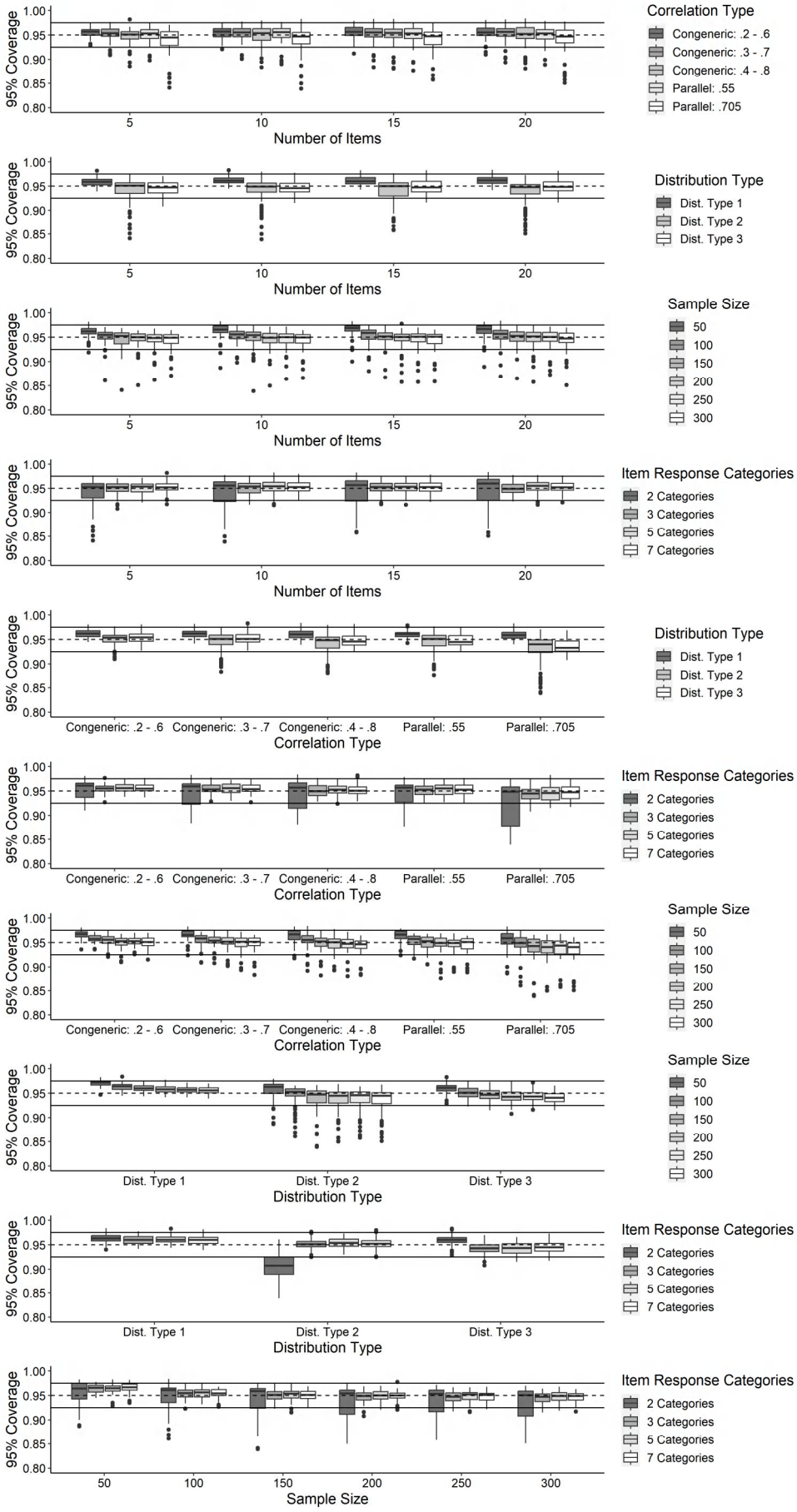

Figure 3. Percentile 95% credible interval coverage for all pairwise simulation conditions. Note. Bayesian alpha normal theory credible interval based on 2000 posterior draws; dashed line at 0.95 and solid lines at acceptable coverage of [0.925, 0.975]. Correlation types indicate structure and range of loadings. Dist. = Distribution.

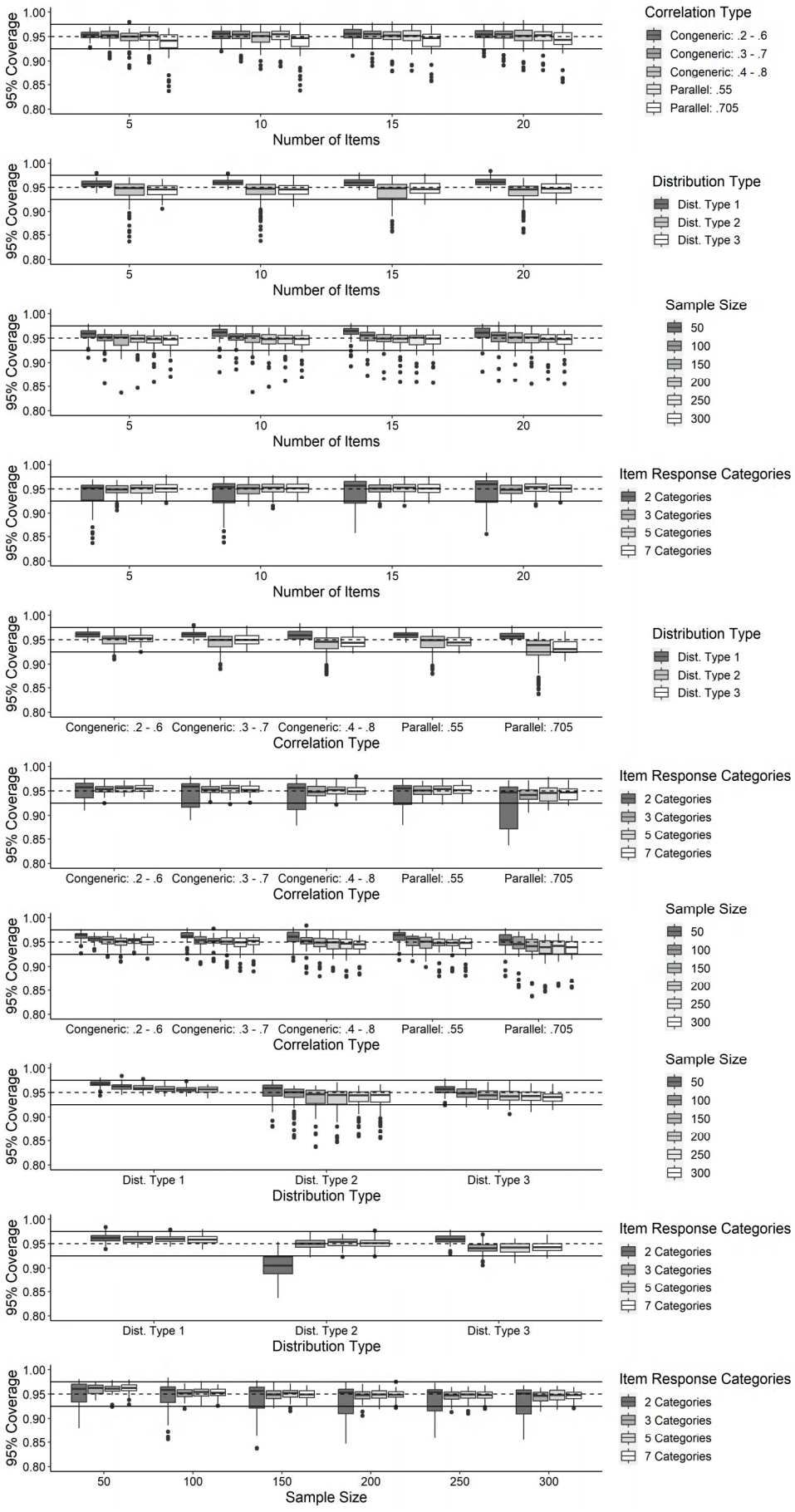

Figure 4. Normal theory 95% credible interval coverage for all pairwise simulation conditions. Note. Bayesian alpha normal theory credible interval based on 2000 posterior draws; dashed line at 0.95 and solid lines at acceptable coverage of [0.925, 0.975]. Correlation types indicate structure and range of loadings. Dist. = Distribution.

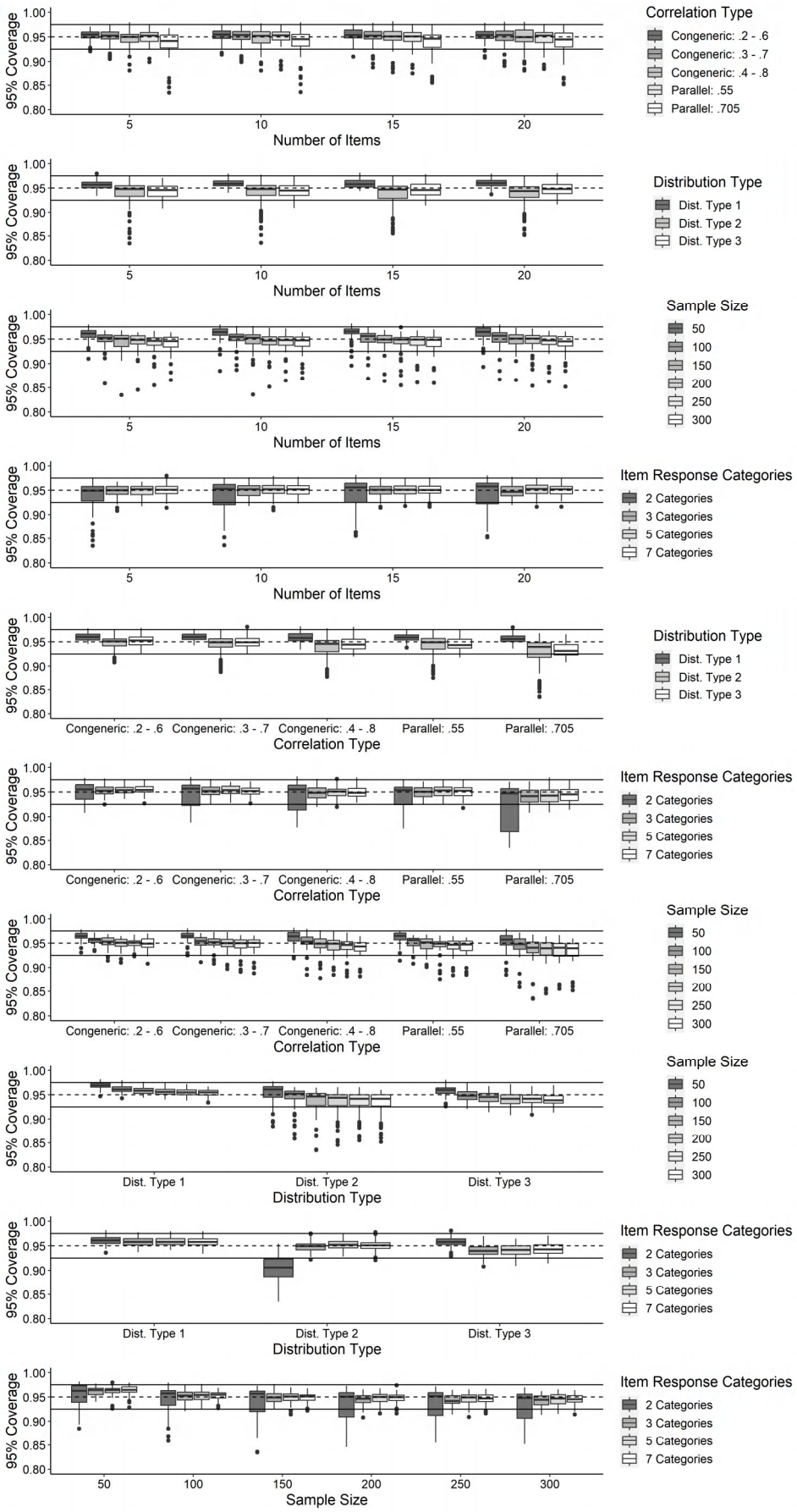

Figure 5. Highest probability density 95% credible interval coverage for all pairwise simulation conditions. Note. Bayesian alpha highest probability density credible interval based on 2000 posterior draws; dashed line at 0.95 and solid lines at acceptable coverage of [0.925, 0.975]. Correlation types indicate structure and range of loadings. Dist. = Distribution.

For the NT CrI (Figure 4), the simulation conditions that involved binary items ($M = 2$), distribution type 2, and/or a parallel model with common loadings of 0.705 tended to have unacceptable coverage. For binary items, performance was impacted across all number of items; for all correlation types except a congeneric model with factor loading of $0.2 - 0.6$; or with a sample size of $150 - 300$. For distribution type 2, performance was impacted across all number of items; or with sample size of $100 - 300$. For a parallel model with common loadings of 0.705, performance was impacted across all number of items; or sample size of $100 - 300$. Like the percentile CrI, the pairwise combinations of binary items with distribution type 2; distribution type 2 with a parallel model with common loadings of 0.705; or binary items with a parallel model with common loadings of 0.705 had the largest performance impact of unacceptable coverage.

For the HPD CrI (Figure 5), the simulation conditions that involved binary items ($M = 2$), distribution type 2, and/or a parallel model with common loadings of 0.705 tended to have unacceptable coverage. For binary items, performance was impacted across all number of items; for all correlation types except a congeneric model with factor loadings of $0.2 - 0.6$; or with a sample size of $150 - 300$. For distribution type 2, performance was impacted across all number of items; or with a sample size of $100 - 300$. For a parallel model with common loadings of 0.705, performance was impacted across all number of items; or with sample size of $100 - 300$. As with the percentile and NT CrIs, the pairwise combinations of binary items with distribution type 2; distribution type 2 with a parallel model with common loadings of 0.705; or binary items with a parallel model with common loadings of 0.705 had the largest performance impact of unacceptable coverage.

In summary, the CrIs investigated were impacted by the same simulation conditions and had similar performance. Specifically, simulation conditions that involved binary items ($M = 2$), distribution type 2, and/or a parallel model with common loadings of 0.705 tended to have unacceptable coverage. For binary items, performance was impacted across all number of items; for all correlation types except a congeneric model with factor loadings of $0.2 - 0.6$; or with a sample size of $150 - 300$. For distribution type 2, performance was impacted across all number of items; or with a sample size of $100 - 300$. For a parallel model with common loadings of 0.705, performance was impacted across all number of items; or with sample size of $100 - 300$. Additionally, the pairwise combinations of binary items with distribution type 2; distribution type 2 with a parallel model with common loadings of 0.705; or binary items with a parallel model with common loadings of 0.705 had the largest performance impact with unacceptable coverage.

3.2. Interval Width

Figure 6 shows the distribution of widths of each CrI method across the main effects of the simulation. Performance amongst the CrIs methods were similar to one another for each simulation condition. For number of items, CrI widths decreased and became less varied as the number items increased. For correlation type, the congeneric model with factor loadings of $0.2 - 0.6$ had the greatest CrI widths and variability. However, the parallel model with a factor loading of 0.705 had the smallest CrI widths and variability. For distribution type, all three distribution types performed similarly with the type I distribution having slightly smaller widths than the other distribution types. For sample size, CrI widths decreased and became less varied as the sample size increased. For item responses categories, CrI width decreased as the number of item responses categories increased.

3.3. Application Example

For demonstration purposes, the coefficient alpha CrIs investigated in the simulation were also applied to real data. This data was pulled from 222 respondents from the Short Form Big Five Inventory-2 (BFI-2-S). The BFI-2-S is a self-report measurement instrument used to assess a respondent's measurement on the constructs of (a) agreeableness, (b) open-mindedness, (c) conscientiousness, (d) negative emotionality, and (e) extraversion [23]. In terms of format, the BFI-2-S consists of 30 five-point Likert-type items where each construct is represented by a set of six items. For this example, suppose a researcher is interested in how agreeableness predicts marriage status. From the data, a coefficient alpha for agreeableness is found to be 0.57. Investigating the literature, the researcher finds that a coefficient alpha for agreeableness was previously estimated by Hahn et al. [24] on a similar measurement instrument, the Short Form Big Five Inventory (BFI-S; [25]). The interest here is to apply information from the previous Hahn et al. study to the data in the current example.

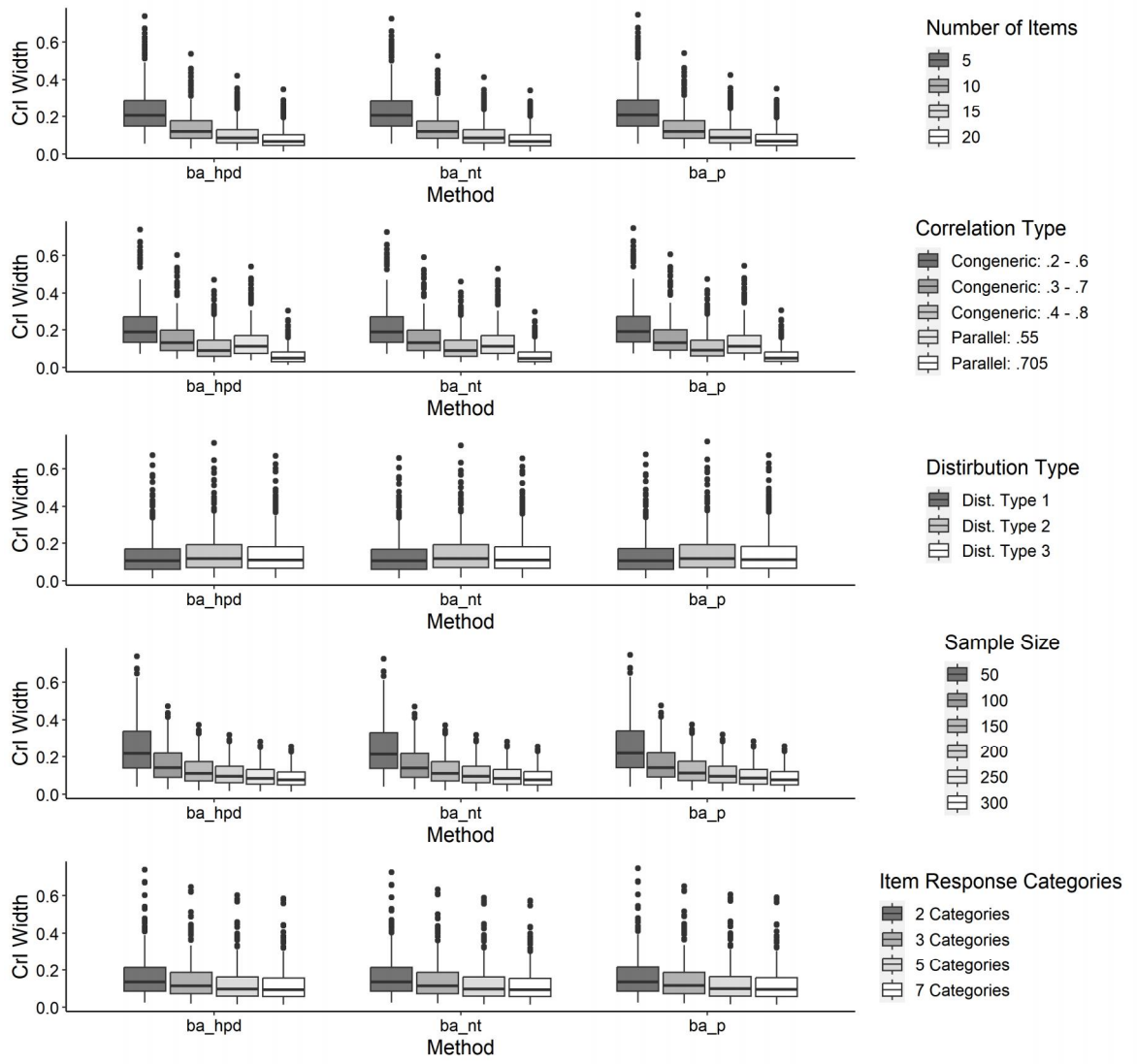

Figure 6. Distribution of the 95% credible interval widths for all main effect simulation conditions. Note. Credible interval types denoted by the following: Bayesian alpha highest probability density (ba_hpd), Bayesian alpha normal theory (ba_nt), Bayesian alpha percentile (ba_p). Bayesian method based on 2000 posterior draws. Correlation types indicate structure and range of loadings. Dist. = Distribution.

From Hahn et al. [24] the coefficient alpha for agreeableness from the BFI-S is 0.44 with a sample size of 598. These can be used as the priors for coefficient alpha and sample size, respectively. A prior variance for coefficient alpha is not directly given here, but an accommodation for this given by DelosReyes and Padilla [18] can be applied here if item variances are available and if it is assumed that the items follow a parallel measurement model. Item variances are given in the study and, for the purposes of demonstration, we assume these items follow a parallel measurement model. To summarize what the accommodation does: item variances are used to calculate a compound symmetric covariance matrix needed to obtain a prior variance for coefficient alpha. The descriptive statistics of the BFI-S for agreeableness are summarized on Table 3 and include item variances. Using the accommodation, the prior variance for coefficient alpha is 0.9408. The results of the 95% coefficient alpha CrIs where priors were used and not used are summarized in Table 4. Bayesian estimation without the use of priors (e.g., non-informative priors) simply results in estimation driven by data alone. Note in Table 4 how the CrIs adjust once priors are used in the estimation: the point estimate is adjusted towards the prior and the intervals widths are reduced.

Table 3. Descriptive statistics for agreeableness from the short form big five inventory (n = 598; Hahn et al. [24]).

Item	Mean	SD	Variance
Is sometimes somewhat rude to others (R)	3.42	1.72	2.96
Has a forgiving nature	5.42	1.36	1.85
Is considerate and kind to others	5.96	1.02	1.04

Note. (R) indicates item was reverse coded.

Table 4. Coefficient alpha 95% credible intervals for agreeableness in the short form big five inventory-2.

Method	Estimate	Lower Bound	Upper Bound	Width
Percentile	0.57	0.49	0.65	0.16
Normal Theory	0.57	0.49	0.65	0.16
High Probability Density	0.57	0.49	0.65	0.16
Percentile with Priors	0.48	0.42	0.54	0.12
Normal Theory with Priors	0.48	0.42	0.54	0.12
High Probability Density with Priors	0.48	0.42	0.54	0.12

Note. Credible intervals based on 2000 posterior draws.

4. Discussion

An alternative method to obtain a Bayesian estimate of coefficient alpha based on a normal posterior was recently introduced and assessed [18]. That initial proposal showed promising results as the CrIs investigated showed generally good results even when items were not parallel (i.e., compound symmetric). However, the method assumes items are normally distributed. Therefore, it is of interest to assess how the alternative method performs when using non-normal items.

Expanding on that previous research, the effect non-normal items had on the alternative Bayesian estimate of coefficient was assessed via simulations. Generally, the CrIs investigated performed well and similarly to one another. However, it was found that simulation conditions that involved binary items, distribution type 2, and/or a parallel model with common loadings of 0.705 tended to have negatively impacted performance. This effect was further pronounced if these simulation conditions were paired with one another. Even so, the results are limited to the conditions investigated in the current study.

A common feature of these conditions is that they have some form of range restriction. First, the alternative method CrIs are based on making direct draws from a normal posterior distribution. However, binary items cannot achieve normality and are thus at odds with draws from the normal posterior distribution. This can be seen as binary items having a range restriction that cannot accurately reflect the variability in a normal distribution. Second, items with skewed distributions tend to have range restrictions because of floor/ceiling effects. This can be seen for the items with skewness greater or equal to 0.97 from Figure 1. Lastly, a parallel model with a common loading of 0.705 has a corresponding compound symmetric item correlation structure with $\rho = .56$. This created the strongest coefficient alphas with an average coefficient alpha of 0.88 (see Table 1). This created a range restriction at the upper limit of 1 for coefficient alpha. Any one of these conditions negatively impacts the CrIs, but any combination of them exacerbated that impact. Range restrictions attenuate variance and accurate variance is necessary to have accurate CrIs.

There are a few directions to expand on the current study. First, it would be of interest to directly compare the performance of the alternative method for obtaining a Bayesian estimate of coefficient alpha by DelosReyes and Padilla [18] to other methods. These comparisons were not done here as the focus was to evaluate the viability of the method on non-normal data. Second, given how the CrIs were impacted by binary items, it may be of interest to explore using a Gibbs sampler when making posterior draws [26]. Unlike the MCMC used here that directly makes draws for a normal posterior, an advantage of the Gibbs sample is that it cycles through data at each draw (i.e., iteration). This cycling though the data allows the data to have more impact on the posterior, which in turn can help improve the situation with binary items. Incidentally, cycling through the data at each draw may also help with skewed items (i.e., non-normal items). Alternatively, given that the tetrachoric correlation is for binary variables, it may be of interest to investigate the tetrachoric correlation as part of the Bayesian estimate of coefficient alpha. Third, it would be of interest to consider exploring the viability of the method with alternative forms of data. The current study explored how the method works in simulation with Likert-type (or ordinal) and binary items with target distributions that were non-normal. These situations were a focus here as they are frequently encountered in behavioral/social science research. In line with this, one direction to consider is how the method works with multilevel/nested data as these situations are also frequently encountered in behavioral/social science research and further developments in estimating reliability in this direction have been making progress [27].

All things considered, the simulation demonstrated that the alternative method to obtain a Bayesian estimate of coefficient alpha was found to be generally viable for non-normal items. There was a negative performance impact if items were binary, came from a distribution type 2, and/or a parallel model with common loadings of 0.705. Conversely, it was shown that the method is viable if items have at least three item response categories, have a distribution less kurtotic than distribution type 2 (i.e., $\text{kurtosis} < 0.88$), and have more variability than that

associated with the parallel model with common loadings of 0.705. Essentially, the method works for non-normal items if the items are allowed to have a healthy amount of variability.

Author Contributions

J.M.V.D.: conceptualization, methodology, investigation, software, writing—original draft preparation; M.A.P.: conceptualization, methodology, supervision, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Application example data are available upon request.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

1. Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* **1951**, *16*, 297–334. https://doi.org/10.1007/BF02310555.
2. Padilla, M.A. A primer on reliability via coefficient alpha and omega. *Arch. Psychol.* **2019**, *3*, 1–15. https://doi.org/10.31296/aop.v3i8.125.
3. Furr, R.M.; Bacharach, V.R. *Psychometrics: An Introduction*, 2nd ed.; Sage: Los Angeles, CA, USA, 2014.
4. Novick, M.R. Coefficient alpha and the reliability of composite measurements. *Psychometrika* **1967**, *32*, 1–13. https://doi.org/10.1007/BF02289400.
5. Beaulieu-Prevost, D. Confidence intervals: From tests of statistical significance to confidence intervals, range hypotheses, and substantial effects. *Tutor. Quant. Methods Psychol.* **2006**, *2*, 11–19. https://doi.org/10.20982/tqmp.02.1.p011.
6. Benjamin, D.J.; Berger, J.O.; Johannesson, M.; et al. Redefine statistical significance. *Nat. Hum. Behav.* **2018**, *2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z.
7. Wilkinson, L.; Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *Am. Psychol.* **1999**, *54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594.
8. Feldt, L.S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika* **1965**, *30*, 357–370. https://doi.org/10.1007/BF02289499.
9. Barchard, K.A.; Hakstian, A.R. The robustness of confidence intervals for coefficient alpha under violation of the assumption of essential parallelism. *Multivar. Behav. Res.* **1997**, *32*, 169–191. https://doi.org/10.1207/s15327906mbr3202_4.
10. Duhachek, A.; Iacobucci, D. Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *J. Appl. Psychol.* **2004**, *89*, 792–808. https://doi.org/10.1037/0021-9010.89.5.792.
11. van Zyl, J.M.; Neudecker, H.; Nel, D.G. On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika* **2000**, *65*, 271–280. https://doi.org/10.1007/bf02296146.
12. Yuan, K.H.; Guarnaccia, C.A.; Hayslip, B., Jr. A study of the distribution of sample coefficient alpha with the Hopkins Symptom Checklist: Bootstrap versus asymptotics. *Educ. Psychol. Meas.* **2003**, *63*, 5–23. https://doi.org/10.1177/0013164402239314.
13. Derogatis, L.R.; Lipman, R.S.; Rickels, K.; et al. The Hopkins symptom checklist (HSCL): A self-report symptom inventory. *Behav. Sci.* **1974**, *19*, 1–15. https://doi.org/10.1002/bs.3830190102.
14. Maydeu-Olivares, A.; Coffman, D.L.; Hartmann, W.M. Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychol. Methods* **2007**, *12*, 157–176. https://doi.org/10.1037/1082-989X.12.2.157.

15. Romano, J.L.; Kromrey, J.D.; Hibbard, S.T. A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educ. Psychol. Meas.* **2010**, *70*, 376–393. https://doi.org/10.1177/0013164409355690.
16. Padilla, M.A.; Divers, J.; Newton, M. Coefficient alpha bootstrap confidence interval under nonnormality. *Appl. Psychol. Meas.* **2012**, *36*, 331–348. https://doi.org/10.1177/0146621612445470.
17. Forsyth, D. *Probability and Statistics for Computer Science*, 1st ed.; Springer: Cham, Switzerland, 2018.
18. DelosReyes, J.M.V.; Padilla, M.A. Obtaining a Bayesian estimate of coefficient alpha using a posterior normal distribution. *Educ. Psychol. Meas.* **2025**, *85*, 829–852. https://doi.org/10.1177/00131644241311877.
19. Lee, P.M. *Bayesian Statistics: An Introduction*, 3rd ed.; Arnold: London, UK, 2004.
20. Padilla, M.A.; Zhang, G. Estimating internal consistency using Bayesian methods. *J. Mod. Appl. Stat. Methods* **2011**, *10*, 277–286. https://doi.org/10.22237/jmasm/1304223840.
21. Bradley, J.V. Robustness? *Br. J. Math. Stat. Psychol.* **1978**, *31*, 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.
22. Anderson, A.A. Assessing statistical results: Magnitude, precision, and model uncertainty. *Am. Stat.* **2019**, *73*, 118–121. https://doi.org/10.1080/00031305.2018.1537889.
23. Soto, C.J.; John, O.P. Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *J. Res. Personal.* **2017**, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004.
24. Hahn, E.; Gottschling, J.; Spinath, F. Short measurements of personality—Validity and reliability of the GSOEP big five Inventory (BFI-S). *J. Res. Personal.* **2012**, *46*, 355–359. https://doi.org/10.1016/j.jrp.2012.03.008.
25. Gerlitz, J.; Schupp, J. *Zur Erhebung der Big-Five-Basierten Persönlichkeitsmerkmale im SOEP [Assessment of Big Five Personality Characteristics in the SOEP]*; DIW: Berlin, Germany, 2005.
26. Hoff, P.D. *A First Course in Bayesian Statistical Methods*; Springer: New York, NY, USA, 2009.
27. Lai, M.H.C. Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychol. Methods* **2021**, *26*, 90–102. https://doi.org/10.1037/met0000287.