2023 ■ Vol. 19 ■ no. 2

Confidence Intervals for the coefficient alpha difference from two independent samples (groups)

Miguel A. Padilla [a]

[a] Old Dominion University

Abstract ■ Four different bootstrap methods for estimating confidence intervals (CIs) for a coefficient alpha difference from two independent samples (groups) were examined. These four CIs were compared to the most promising non-bootstrap CI alternatives in the literature. All CIs were assessed with a Monte Carlo simulation with conditions similar to previous research. The results indicate that there is a clear order in coverage performance of the CIs. The bootstrapped highest density interval had the best coverage performance across all simulation conditions. Yet, it was impacted by unequal sample sizes when one of the groups had the smallest sample size investigated of 50, or when items came from a compound symmetric correlation matrix with $\rho = 0.64$. Regardless of the simulation condition, the percentile bootstrap is a good alternative as long as both group sample sizes were 200 or more.

Keywords ■ Reliability; internal consistency; independent samples; bootstrap. **Tools** ■ R.

✉ mapadill@odu.edu

10.20982/tqmp.19.2.p194

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers ■ One anonymous reviewer

Introduction

Behavioral/ social scientist are often interested in the reliability for a set of items from a test, inventory, questionnaire, or some other measurement instrument for a construct. One form of reliability is internal consistency via coefficient alpha (α_c), which was proposed by Cronbach (1951) based on the work of Guttman (1945). Since then, coefficient alpha has gone on to become the most popular measure of reliability in the sciences (Bollen, 1989, p. 215, Hogan et al., 2000). Its popularity stems from three features. First, coefficient alpha is computationally simple, requiring only the item covariance matrix. Second, coefficient alpha is appropriate for continuous, ordinal, or binary items. Third, coefficient alpha only requires one administration of the measurement instrument. Even so, until recently, inference research about coefficient alpha from one sample has been sporadic, but even more sporadic for the coefficient alpha difference from two independent samples (groups). In this respect, confidence interval (CI) research for coefficient alpha from one sample offers a basis for the development of the coefficient alpha difference from two independent samples (groups).

Feldt (1965) first derived the sampling distribution and corresponding CI for coefficient alpha. Feldt's derivations assumed the items to be normally distributed and parallel (Lord et al., 1968). The parallel assumption implies that the item covariance matrix is compound symmetric (Padilla, 2019). Unfortunately, the CI is not valid if the parallel assumption is violated (Barchard & Hakstian, 1997). This may be a reason why this coefficient alpha CI is not widely used.

van Zyl et al. (2000) showed that the original estimate of coefficient alpha is a maximum likelihood (ML) estimate that has an asymptotic normal sampling distribution. As such, the original estimate of coefficient alpha is based on normal theory (NT) for the sampling distribution. Their results assume the items to be normally distributed with a positive definite item covariance matrix and make no assumption about the structure of the item covariance matrix. This provided a foundation for inference about coefficient alpha. Based on this foundation, Duhachek and Iacobucci (2004) proposed a coefficient alpha NT CI. In the same work via simulation, the authors compared the coefficient alpha NT CI to other CIs that included the one derived by Feldt (1965). The results suggested that the NT CI coverage outperformed all the investigated CIs across all simulation conditions. The simulation was based on α_c from a compound symmetric and unstructured item covariance matrix with 5 and 7 multivariate normal items and sample sizes ranging from 30 to 200.

Other coefficient alpha CIs have been proposed in the literature (Bonett, 2002, 2003, 2010; Bonett & Wright, 2015). The accumulation of this research resulted in two one-sample coefficient alpha CIs (BTT1b and BTT1c) along with an independent samples (groups; BTT2b) coefficient alpha CI. The BTT1c and BTT2b both use the NT variance (van Zyl et al., 2000). In a small simulation, Bonett (2010) demonstrated that the BTT1b CI, had good coverage with small samples. The simulation was based on α_c from a compound symmetric item covariance matrix with 2 to 18 multivariate normal items and sample sizes ranging from 25 to 100. In another small simulation study, Bonett and Wright (2015) demonstrated that the BTT1c CI coverage performed slightly better than the one sample NT CI. The simulation was based on α_c from an autoregressive of order 1, AR(1), item covariance matrix with 4 multivariate normal items and sample sizes ranging from 10 to 200.

More recently, Padilla et al. (2012) examined three standard bootstrap CIs for coefficient alpha: the normal theory bootstrap (NTB), percentile bootstrap (PB), and bias corrected and accelerated (BCa). In the same study, the authors compared these three bootstrap CIs to the following non-bootstrap CIs: the BTT1b and NT CIs, among others. In the study, it was concluded that the NTB CI had the best coverage across all conditions investigated with only 4 instances of unacceptable coverage, followed by the PB and BCa CIs. The simulation included α_c based on compound symmetric and unstructured item correlations with 5 to 20 items and sample sizes ranging from 50 to 300. However, if computational power (or time) is an issue and items were normally distributed or had little skew, the BTT1b followed by the NT CI are good alternatives.

In summary, the research discussed thus far mainly focused on a one-sample coefficient alpha CI. Note that the one-sample coefficient alpha is the original coefficient alpha proposed by Cronbach (1951). In this respect, there appears to be at least three promising forms of coefficient alpha CIs: the NT, BTT1b, and the bootstrap. Although these studies mentioned how to extend each of these one-sample CIs to two independent samples (groups), the latter were not further investigated.

Researchers may have interest in comparing the reliability, via coefficient alpha, of a measurement instrument from two independent samples (groups). In such a situation, researchers would like to make an inference about coefficient alpha being different between the two samples (groups); e.g., coefficient alpha being stronger (weaker) for one sample (group) than the other. Some examples where reliability differences may be of interest are between males and females, treatment and no treatment, taking a test with or without note cards, etc.

The focus of the present study is to investigate the performance of CIs for the coefficient alpha difference from two independent samples (groups) via simulation. For comparison purposes and to provide guidelines between the CIs, similar simulation conditions from previous studies were used. The following CI estimates were investigated: NT, BTT2b, and the bootstrap.

Coefficient Alpha Difference

Suppose there is a set of k items that measure a single attribute/ construct. It is common to estimate the reliability for the sum or composite of the items:

$$x = \sum_{j=1}^{k} x_j. \tag{1}$$

Coefficient alpha is a popular choice for estimating this form of reliability, and is defined as

$$\alpha_c = \frac{k}{k-1}\left[1 - \frac{tr(\mathbf{\Sigma})}{\mathbf{1}'\mathbf{\Sigma}\mathbf{1}}\right] \tag{2}$$

where $tr(\cdot)$ is the trace operator for a matrix, $\mathbf{\Sigma}$ is the $k \times k$ item covariance matrix, and $\mathbf{1}$ is a $k \times 1$ vector of ones. Using the sample estimate $\hat{\mathbf{\Sigma}}$ in place of $\mathbf{\Sigma}$ in Equation 2 gives the coefficient alpha estimate $(\hat{\alpha}_c)$.

Now consider two independent samples (groups) with estimated $\hat{\alpha}_{c1}$ of size n_1 with k_1 items and $\hat{\alpha}_{c2}$ of size n_2 with k_2 items. Here, a hypothesis of interest is $\mathcal{H}_0 : \ \alpha_d = 0$, where $\alpha_d = \alpha_{c1} - \alpha_{c2}$. A CI for this hypothesis can be obtained from non-bootstrapped or bootstrapped methods.

Non-bootstrapped Coefficient Alpha Difference CIs

van Zyl et al. (2000) showed that the coefficient alpha estimate $(\hat{\alpha}_c)$ is asymptotically distributed as

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

where the variance is

$$\sigma_c^2 = \frac{2k^2}{(k-1)^2} \times \left(\frac{(\mathbf{1}'\mathbf{\Sigma}\mathbf{1})\left(tr\left(\mathbf{\Sigma}^2\right) + tr(\mathbf{\Sigma})^2\right) - 2tr\left(\mathbf{\Sigma}\right)\left(\mathbf{1}'\mathbf{\Sigma}^2\mathbf{1}\right)}{\left(\mathbf{1}'\mathbf{\Sigma}\mathbf{1}\right)^3}\right). \tag{4}$$

Using the sample estimate $\hat{\mathbf{\Sigma}}$ in place of $\mathbf{\Sigma}$ in Equation 4 gives the coefficient alpha variance estimate $\left(\hat{\sigma}_c^2\right)$. Therefore, the standard error (SE) is

$$SE\left(\hat{\alpha}_c\right) = \sqrt{\frac{\hat{\sigma}_c^2}{n}} \tag{5}$$

and a CI for the one-sample coefficient alpha estimate is formed as

$$\hat{\alpha}_c \pm z_{\alpha/2} SE\left(\hat{\alpha}_c\right) \quad (6)$$

where $z_{\alpha/2}$ is a standard normal variate with the desired α level.

Let α_{c1} and α_{c2} be coefficient alphas for group1 and group 2, respectively. It then follows that a $\alpha_d = \alpha_{c1} - \alpha_{c2}$ CI for the two independent groups (samples) is formed as

$$\hat{\alpha}_d \pm z_{\alpha/2} SE\left(\hat{\alpha}_d\right), \quad (7)$$

where

$$SE\left(\hat{\alpha}_d\right) = \sqrt{\frac{\hat{\sigma}_{c1}^2}{n_1} + \frac{\hat{\sigma}_{c2}^2}{n_2}}. \quad (8)$$

Bonett and Wright (BTT1c; 2015) proposed a CI for the one-sample coefficient alpha estimate based on the transformation below

$$\hat{z}_c = \ln\left(1 - \hat{\alpha}_c\right) - b \quad (9)$$

where $b = \ln\left(n/(n-1)\right)$ is a bias correction. The CI for $\exp\left(\hat{z}_c\right)$ is then formed as

$$1 - \exp\left[\hat{z}_c \pm z_{\alpha/2} SE\left(\hat{z}_c\right)\right], \quad (10)$$

where

$$SE\left(\hat{z}_c\right) = \sqrt{\frac{\hat{\sigma}_c^2}{(n-3)\left(1-\hat{\alpha}_c\right)^2}}. \quad (11)$$

Based on results from Zou (2007), Bonett and Wright (BTT2b; 2015) also proposed a CI for α_d from two independent samples (groups) as

$$\begin{aligned} L &= \hat{\alpha}_d - \sqrt{\left(\hat{\alpha}_{c1} - L_1\right)^2 + \left(\hat{\alpha}_{c2} - U_2\right)^2} \\ U &= \hat{\alpha}_d + \sqrt{\left(\hat{\alpha}_{c1} - U_1\right)^2 + \left(\hat{\alpha}_{c2} - L_2\right)^2} \end{aligned} \quad (12)$$

where $\hat{\alpha}_{c1}$, lower bound L1, and upper bound U1, are the group 1 estimates, and $\hat{\alpha}_{c2}$, lower bound L2, and upper bound U2 are the group 2 estimates from Equation 10.

Bootstrapped Coefficient Alpha Difference CIs

The bootstrap for the coefficient alpha difference from two independent samples (groups) can be summarized in three steps. Let $\mathbf{X}_1 = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}\right)$ be the group 1 sample where each $\mathbf{x}_1$ is a $1 \times k_1$ vector, and $\mathbf{X}_2 = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_2}\right)$ be the group 2 sample where each $\mathbf{x}_2$ is a $1 \times k_2$ vector. First, independently obtain a bootstrap sample for each group $\mathbf{X}_1^{(b)} = \left(\mathbf{x}_1^{(b)}, \mathbf{x}_2^{(b)}, \ldots, \mathbf{x}_{n_1}^{(b)}\right)$ and $\mathbf{X}_2^{(b)} = \left(\mathbf{x}_1^{(b)}, \mathbf{x}_2^{(b)}, \ldots, \mathbf{x}_{n_2}^{(b)}\right)$ in which b is a random resample with replacement from $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Note that, $\mathbf{X}_1^{(b)}$and $\mathbf{X}_2^{(b)}$ have the same sample sizes as $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Second, compute the b^{th} bootstrap estimate of the coefficient alpha difference $\left(\hat{\alpha}_d^{(b)}\right)$ from $\mathbf{X}_1^{(b)}$ and $\mathbf{X}_2^{(b)}$. Lastly, $\hat{\alpha}_d^{(1)}, \hat{\alpha}_d^{(2)}, \ldots, \hat{\alpha}_d^{(B)}$ is the empirical sampling distribution (ESD) for $\hat{\alpha}_d$ with $b = 1, 2, \ldots, B$ bootstrap samples. The ESD can then be summarized for statistical inference about $\hat{\alpha}_d$.

The bootstrap SE estimate is

$$SE\left(\hat{\alpha}_d\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\alpha}_d^{(b)} - \bar{\alpha}_d\right)^2} \quad (13)$$

where $\hat{\alpha}_d^{(b)}$ is the estimated coefficient alpha difference from the bth bootstrap replicate and

$$\bar{\alpha}_d = \frac{1}{B}\sum_{b=1}^{B}\hat{\alpha}_d^{(b)}. \quad (14)$$

Four CIs for the bootstrap were investigated. First, the normal theory bootstrap (NTB) CI is computed as $\hat{\alpha}_d \pm z_{\alpha/2} SE\left(\hat{\alpha}_d\right)$. Second, the percentile bootstrap (PB) CI is obtained by computing the $\alpha/2$ and $1 - \alpha/2$ percentiles from the $\hat{\alpha}_d$ ESD at the desired α level. Third, the bias corrected and accelerate (BCa) CI is the PB CI that is adjusted in two ways: 1) corrects for bias, and 2) corrects for skewness (or acceleration). The fourth CI is based on the highest probability density interval (Chen & Shao, 1999). This kind of interval is the shortest and captures the specified probability in the mass of any distribution. Here, the interval is being used on the ESD generated from the bootstrap samples and hence why it is being referred to as the bootstrapped highest density interval (BHDI). Note that the NTB CI assumes the ESD to be normally distributed, whereas the PB, BCa, and BHDI make no assumption about the shape of the ESD. For details about the first three bootstrap CIs, see Efron and Tibshirani (1998), and for the BHDI, see Chen and Shao.

Method

Simulation Design

There were two Monte Carlo simulations. Simulation 1 consisted of the following design: 4 (# of items) × 4 (item correlation structure) × 4 (item response categories) × 21 (pairwise sample). Simulation 2 had the same conditions but only consisted of 2 items and was investigated as a special case for where only two items are available. All simulated items were Likert or binary. For each simulation condition, 10,000 replications were obtained. Below are the specific simulation conditions investigated.

Number of Items (k). For simulation 1, and to make the results comparable to previous research, the following number of items were used: $k = 5, 10, 15, 20$ (Maydeu-Olivares et al., 2007; Padilla et al., 2012). Simulation 2 only had two items ($k = 2$).

Item Correlation Structure ($\mathbf{P}$). For simulation 1, four different item correlations matrices for $\mathbf{P}$ were investigated. The first three correlation matrices are from parallel items with a one factor model containing factor loadings of $\lambda_1 = 0.4$, $\lambda_2 = 0.6$, and $\lambda_3 = 0.8$ corresponding to compound symmetric correlation matrices (CSCM) with $\rho_1 = 0.16$, $\rho_2 = 0.36$, and$\rho_3 = 0.64$, respectively. The fourth correlation matrix was from congeneric items with a one factor model containing the loadings of $\lambda_4 = 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7$ corresponding to an unstructured correlation matrix (UCM). These correlation structures were investigated by Maydeu-Olivares et al. (2007), but here the congeneric structure was modified for five items instead of seven. For simulation 2, the $k = 2$ congeneric items had factor loadings of $\lambda_4 = 0.3,\ 0.7$ for the UCM; i.e., the smallest and largest factor loadings from the above congeneric items.

Item Response Categories (M). The following item response categories were investigated: 2, 3, 5, 7 (Maydeu-Olivares et al., 2007; Padilla et al., 2012). For all items with $M > 2$, $\boldsymbol{\nu}$ was chosen so that the resultant categorized items had skewness = kurtosis = 0. For items with $M = 2$ (binary), $\boldsymbol{\nu}$ was chosen so that the resultant categorized items had skewness = 0 and kurtosis = -2 (excess kurtosis). For simulation , the combination of k and $\mathbf{P}$ created a coefficient alpha α_c ranging from 0.36 to 0.97 in each group. For simulation 2, the combination of k and $\mathbf{P}$ created a coefficient alpha α_c ranging between 0.19 and 0.75 in each group.

Pairwise Sample Size (q). For each of the two groups in both simulations, the following sample sizes were investigated: 50, 100, 150, 200, 250, 300. These sample sizes were like those used in Padilla and Divers (2013). The unique pairwise sample sizes were determined by first letting $q = 6$ be the number of sample sizes investigated. Then the total number of unique pairwise sample sizes investigated was $q(q+1)/2 = 21$.

The simulation is outlined below:

1. Let $g = 1,\ 2$ index the groups. Based on the factor model loadings (λ), select the correlation structures for the $k_1 \times k_1$ matrix $\mathbf{P}_1$ and $k_2 \times k_2$ matrix $\mathbf{P}_2$, where k_g is the number of items and $\mathbf{P}_g$ the item correlation matrix for group g. Note that $\mathbf{P}_1 = \mathbf{P}_2$ and $k_1 = k_2$, but subscripts are being used to indicate the there are two separate groups.
2. Select a set of $\boldsymbol{\nu}$ thresholds to categorize items.
3. Generate multivariate standard (unit) normal data as a $n_1 \times k_1$ matrix $\mathbf{Z}_1 \sim N(0,\ \mathbf{P}_1)$ and a $n_2 \times k_2$ matrix $\mathbf{Z}_2 \sim N(0,\ \mathbf{P}_2)$ with sample sizes n_1 and n_2, respectively.
4. Categorize z_1 in $\mathbf{Z}_1$ and z_2 in $\mathbf{Z}_2$ with the $\boldsymbol{\nu}$ thresholds into x_1 in $\mathbf{X}_1$ and x_2 in $\mathbf{X}_2$, as follows: $x_g = m$ if $\nu_m < z_g \leq \nu_{m+1}$ for group $g = 1,\ 2$ and $m = 0, 1, ..., M-1$, where $\nu_0 = -\infty$ and $\nu_M = \infty$, and M is the item response categories. For example, when $M = 3$ for $g = 1$ (group 1): $x_1 = 0$ if $\nu_0 = -\infty < z_1 \leq \nu_1 = -.97$; $x_1 = 1$ if $\nu_1 = -.97 < z_1 \leq \nu_2 = 0.97$; $x_1 = 2$ if $\nu_2 = 0.97 < z_1 < \nu_3 = \infty$.
5. Estimate the $\hat{\alpha}_d$ CI from $\mathbf{X}_1$ and $\mathbf{X}_2$.
6. Compute α_d from $\mathbf{P}_1$, $\mathbf{P}_2$, and the $\boldsymbol{\nu}$ thresholds. See Maydeu-Olivares et al. (2007) for full details.
7. Determine if the $\hat{\alpha}_d$ CI contain α_d.

Evaluating CIs

In each simulation replication, $\hat{\alpha}_d$ and corresponding quantities were estimated and evaluated. All $100(1-\alpha)\%$ CIs for $\hat{\alpha}_d$ were estimated with $\alpha = 0.05$. For the bootstrap methods, 2,000 bootstrap samples were used. CI coverage is defined as the proportion of estimated CIs that contain α_d. The proportion was evaluated with Bradley's (1978) stringent criterion, defined as $1 - 1.1\alpha \leq 1 - \alpha* \leq 1 - 0.9\alpha$ where α* is the true Type I error probability. Therefore, acceptable coverage is given by $[0.945, 0.955]$. Additionally, coverage symmetry was also evaluated. Here, the lower asymmetry is defined as the proportion of times the upper CI limit is below α_d, whereas the upper asymmetry is defined as the proportion of times the lower CI limit is above α_d. It is possible for CI estimates to have the same coverage with different coverage symmetry. Lastly, CI width was also evaluated as it provides precision information about a CI estimate. It is possible for CI estimates to have the same coverage with different CI widths (i.e., levels of precision). As such, coverage symmetry and CI width are relevant if CI estimates have similar coverage probability.

Results

In terms of coverage, the BHDI had the best performance. The major impact on most of the CIs was a strong compound symmetric item correlation structure and a small sample size. Even so, results are presented in the context of pairwise sample size.

Simulation 1 ($5 \leq k \leq 20$) Non-bootstrapped CIs

The BTT2b CI had decent coverage; see Figure 1. However, it was impacted by some of the simulation conditions. First, the correlation structure had an impact as the BTT2b CI was more variable and conservative ($\alpha < .05$) as the correlation magnitude increased for the CSCM. However, the BTT2b CI got less variable and more liberal ($\alpha > .05$) with an UCM. Second, the item response categories had an impact as the BTT2b CI did not have good coverage with two categories (binary items; $M = 2$). For the remaining item response categories, the BTT2b CI had decent coverage. Third, the number of items had an impact as the BTT2b becomes liberal with 20 items ($k = 20$). For the other number of items, the BTT2b CI had good coverage. Even so, except for the a CSCM = 0.64, $M = 2$, or $k = 20$, coverage was good with an equal sample size between the groups, or when the sample size increases regardless of group sample sizes being equal.

For the NT CI, coverage was unacceptable; see Figure 2. Regardless of the simulation condition, the NT CI tended to be too liberal but it tended to be most impacted by CSCM and binary items.

Simulation 1 ($5 \leq k \leq 20$) Bootstrapped CIs

The NTB CI had a similar pattern of coverage as the NT CI; see Figure 3. Here, coverage was also unacceptable, but was also too liberal.

The PB CI had acceptable coverage when the sample size was 200 or greater regardless of group sample sizes being equal and regardless of the simulation condition; see Figure 4. However, the PB CI had conservative coverage when the sample size was 150 or less regardless of group sample size being equal and regardless of simulation condition. It was particularly conservative when $n_1 = 50$, the smallest sample size investigated, and regardless of the n_2 sample size.

The BCa CI tended to have acceptable coverage; see Figure 5. In general, acceptable coverage occurred when the sample sizes between the groups were equal, or when the sample size increased regardless of group sample sizes being equal. However, when $n_1 = 50$ and $n_2 \geq 150$, the BCa CI became unacceptably conservative. In addition, it tended to be liberal when with a CSCM = 0.64, an UCM, $M = 7$ categories, or $k = 20$ items.

The BHDI had a similar pattern of coverage as the BCa; see Figure 6. However, coverage was acceptable when both group sample sizes were 100 or more. Even so, it was impacted by the item correlation structure and sample size pairing. First, the BHDI tends to be liberal with a CSCM = 0.64. In addition, when $n_1 = 50$ and $n_2 \geq 100$, the BHDI became conservative, and unacceptably conservative when $n_2 \geq 200$.

Simulation 2 ($k = 2$) Non-bootstrapped CIs

BTT2b CI had decent coverage; see Figure 7. However, it was impacted by some of the simulation conditions. First, the correlation structure had an impact as the CI did not have good coverage with CSCM = 0.64. Even though the CI had good coverage for the remaining structures, it does get more conservative as the correlation gets stronger for the CSCM. Second, the item response categories had an impact on the CI as it did not have good coverage with two categories (binary items; $M = 2$). However, for the remaining items response categories ($M > 2$), the CI had good coverage and got better as the item response categories increased. Lastly, for the condition with good coverage, coverage improved with an equal sample size and/or when the sample size increased for either group.

For the NT CI, coverage was unacceptable; see Figure 8. In most conditions, coverage tended to be too liberal. Interestingly, it was noticeably impacted by the same conditions as the BTT2b (i.e., CSCM = 0.64 or $M = 2$).

Simulation 2 ($k = 2$) Bootstrapped CIs

The NTB CI had a similar pattern of liberal unacceptable coverage as the NT CI; see Figure 9. Here, coverage was also unacceptably too liberal.

The PB CI had acceptable coverage when the sample size was 250 or greater regardless of the group sample size being equal and regardless of the simulation condition; see Figure 10. However, the PB CI tended to have conservative coverage when the sample size was 200 or less regardless of the group sample size being equal and regardless of simulation condition. It was particularly unacceptably conservative when $n_1 = 50$, the smallest sample size investigated, and regardless of the n_2 sample size.

The BCa CI tended to have acceptable coverage as long as one of the sample sizes was 250 or greater regardless of group sample size being equal and regardless of simulation condition; see Figure 11. Even so, the BCa CI tended to have conservative coverage across all simulation conditions.

For the BHDI, coverage was acceptable when $n_1 \geq 100$ and $n_2 \geq 100$ regardless of the group sample sizes being equal and regardless of the simulation condition; see Figure 12. As such, it was unacceptably liberal when $n_1 = n_2 = 50$ and became unacceptably conservative when $n_1 = 50$ and $n_2 \geq 200$.

CI Widths

The average CI widths and coverage symmetries are presented in Table 1. For the most part, these results supported those of CI coverage. In general, there was little variability between the CI widths. For simulation 1 $(5 \leq k \leq 20)$, the BTT2b, NT, BCa, and BHDI had equal widths that were also narrowest. The remaining CIs had equal widths that were also the widest. For simulation 2 $(k = 2)$, the NT hand the narrowest width followed by the BHDI. The remaining CIs had equal widths that were also the widest.

CI Coverage Symmetry

For simulation 1 $(5 \leq k \leq 20)$, the BTT2b, BCa, and BHDI had good coverage symmetry. However, the NT NTB, and PB did not have good coverage symmetry. The NT and NTB tended to be asymmetric that lacked coverage in the lower tail probabilities. On the other hand, the PB tended to have to be asymmetric in favor of the lower tail probabilities.

Table 1 ■ Average CI width and coverage symmetry over all simulation conditions

CI method	CI width	CI coverage symmetry		
Simulation 1				
BTT2b	0.18	0.024	0.954	0.022
NT	0.18	0.018	0.957	0.025
NTB	0.19	0.017	0.960	0.023
PB	0.19	0.032	0.945	0.023
BCa	0.18	0.026	0.950	0.024
BHDI	0.18	0.026	0.950	0.024
Simulation 2				
BTT2b	0.61	0.025	0.947	0.028
NT	0.58	0.015	0.953	0.032
NTB	0.61	0.014	0.959	0.027
PB	0.61	0.028	0.944	0.028
BCa	0.61	0.028	0.943	0.029
BHDI	0.60	0.022	0.950	0.028

Note. Bonett (BTT2b), Normal Theory (NT), Normal Theory Bootstrap (NTB), Percentile Bootstrap (PB), Bias-corrected & accelerated (BCa), and Bootstrap Highest Density Interval (BHDI).

For simulation 2 $(k = 2)$, the BTT2b and BHDI had good coverage symmetry. The remaining CIs did not have good coverage symmetry. The NT and NTB tended to be asymmetric in favor of the upper tails. The PB and BCa tended to have symmetric coverage with high lower and upper tail probabilities.

Point Estimate Bias

Given the performance of some of the CIs, parameter estimate bias was investigated. Here, bias is for $\hat{\alpha}_d = \hat{\alpha}_{c1} - \hat{\alpha}_{c2}$ defined as $\hat{\alpha}_{dbias} = (\hat{\alpha}_{c1} - \hat{\alpha}_{c2}) - (\alpha_{c1} - \alpha_{c2})$. Bias for all simulation combinations was inspected, and no bias was visually observed; see Figures 13 and 14. Even so, there was a small impact when $n_1 = 50$ and $n_2 \geq 100$. In this situation, $|\hat{\alpha}_{dbias}| \leq 0.02$ for Simulation 1 $(5 \leq k \leq 20)$ and $|\hat{\alpha}_{dbias}| \leq 0.03$ for Simulation 2 $(k = 2)$. Again, bias was not a concern for the overall study.

Discussion

The performance of CIs for the difference in coefficient alphas from independent samples (groups) was investigated under several simulation conditions. Coefficient alpha is a reliability index for a composite of at least (essentially) tau equivalent items measuring one dimension, and corresponding CI research has mostly focused on coefficient alpha from one sample (Bonett & Wright, 2015; Padilla et al., 2012). To date, these CIs have not been investigated in a simulation design as done here. In general, the nonparametric bootstrap CI methods (PB, BCa, and BHDI) had the best coverage with some distinctions. Even so, all CIs were impacted by sample size and correlation structure.

In general, the BHDI and BCa CIs were impacted by unequal samples sizes when one of the groups was at the smallest sample size investigated ($n_1 = 50$). The BHDI and BCa CI tended to get conservative as the second sample size increased above 50. This occurred regardless of the other simulation conditions. However, under the compound symmetric correlation structure with $\rho = 0.64$, the BHDI and BCa CI tended to get liberal as the sample size for both groups got larger. Except for the condition with two items ($k = 2$), the BCa CI had virtually the same coverage pattern as the BHDI but was slightly less optimal. However, with only two items, the BCa CI did not have adequate coverage performance. Interestingly, the PB CI had good coverage across all simulation conditions and unequal sample sizes as long as both group sample sizes were 200 or more. Thus, to provide good coverage as nonparametric methods, the BHDI, BCa, and PB CI need a sample size of at least 100, 100, and 200 for at least one of the groups, respectively.

A compound symmetric correlation structure with $\rho = 0.64$ from parallel items impacted all the CI methods. This condition created range restrictions at the coefficient alpha upper limit of 1 as it had the strongest coefficient alphas that ranged from 0.80 to 0.97 for 5 to 20 items, respectively. It is known that the bootstrap struggles with distributions with extreme data or with highly peaked distributions with range restrictions (Efron & Tibshirani, 1998). Interestingly, this was not an issue with two items for the nonparametric bootstrap CI methods. Here, coefficient alpha ranged from 0.61 to 0.75, respectively. While this is an interesting result, parallel items may not be the most realistic. A more realistic situation is an unstructured correlation matrix from congeneric items (Graham, 2006; Padilla, 2019), and in this situation, the nonparametric bootstrap CI methods performed well under the previously specified conditions. It should also be noted that the bias what not an issue for any of the methods. So, the issue is capturing the ESD to obtain the corresponding SE and/or percentiles.

Within the context of the simulation, there is a clear order of performance between the CIs investigated. The BHDI had the best performance in that it had consistent acceptable coverage under all but the two previously specified conditions. This even included when there were only two items. This was followed by the PB CI in a large sample size situation. Even so, a recommendation can be made. If the sample size is 100 or more for either group, the BHDI is superior. However, if the sample size is 200 or more for either group, the PB CI is a good alternative.

Given the results, more research is warranted. The results here were obtained using Likert items that were normally distributed or binary items that were symmetrically distributed. However, it is unlikely that applied data will follow such distributions. As such, future research should investigate the impact of nonnormal data on at least the nonparametric bootstrap CI methods investigated here.

Through the simulation results, four points can be made about the recommended CI methods. First, all items were Likert or binary (not continuous) and it had little to no impact. Second, an unstructured correlation matrix from congeneric items had no impact on the recommended CI methods. These two results are consistent with past research (Padilla et al., 2012). Third, a sample size of at least 100 for the BHDI or 200 for the PB CI in either group is needed for the recommended CI methods to achieve acceptable coverage. This may seem like a large sample size requirement, but recall that the recommended CI methods are nonparametric, the smallest correlation matrix size was 5×5, and the groups were independent. Each of these situations on their own require a decent sample size. Lastly, interested researchers can obtain the R syntax for the nonparametric bootstrap CI methods in the appendix as well as on the journal's web site.

References

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

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340. http://www.jstor.org/stable/3648121

Bonett, D. G. (2003). Sample size requirements for comparing two alpha coefficients. *Applied Psychological Measurement*, *27*(1), 72–74. doi: 10.1177/0146621602239477.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*(4), 368–385. doi: 10.1037/a0020142.

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. doi: 10.1002/job.1960.

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

Chen, M.-H., & Shao, Q.-M. (1999). Monte carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, *8*(1), 69–92. doi: 10.1080/10618600.1999.10474802.

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

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(5), 792–808. doi: 10.1037/0021-9010.89.5.792.

Efron, B., & Tibshirani, R. (1998). *An introduction to the bootstrap*. CRC Press.

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

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944. doi: 10.1177/0013164406288165.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. doi: 10.1007/bf02288892.

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. doi: 10.1177/00131640021970691.

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

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*(2), 157–176. doi: 10.1037/1082-989x.12.2.157.

Padilla, M. A. (2019). A primer on reliability via coefficient alpha and omega. *Archives of Psychology*, *3*(8), 1–15. https://archivesofpsychology.org/index.php/aop/article/view/125

Padilla, M. A., & Divers, J. (2013). Bootstrap interval estimation of reliability via coefficient omega. *Journal of Modern Applied Statistical Methods*, *12*(1), 78–89. doi: 10.22237/jmasm/1367381520.

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

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

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. doi: 10.1037/1082-989X.12.4.399.

Appendix

```
# load packages
library(boot)
library(coda)

# import data
dat1<- read.csv("~/two_group_alpha.csv")
head(dat1)
##   x1 x2 x3 group
## 1  0  3  2     1
## 2  0  0  0     1
## 3  0  2  0     1
tail(dat1)
##     x1 x2 x3 group
## 195  1  0  0     2
## 196  1  0  0     2
## 197  3  0  1     2

# bootstrap syntax
alphac2<- function(dat, i) {
   j<- ncol(dat)
   k<- j - 1
   datbt<- dat[i,1:k]
   a1<- (k/(k-1))*(1 - sum(diag(cov(datbt[dat[,j]==1,])))/sum(cov(datbt[dat[,j
   ]==1,])))
   a2<- (k/(k-1))*(1 - sum(diag(cov(datbt[dat[,j]==2,])))/sum(cov(datbt[dat[,j
   ]==2,])))
   alDif<- a1 - a2
   return(alDif)
}

set.seed(105)
(btalpc <- boot(data=dat1, statistic=alphac2, strata=dat1$group, R=2000))
btSmpl  <- btalpc$t;

alphCI  <- boot.ci(btalpc, conf=0.95, type=c("perc", "bca"))
print(alphCI)

# BHDI syntax from CODA package
btSmpl  <- as.mcmc(btSmpl)
HPDinterval(btSmpl, prob=0.95)
```

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

Padilla, M. A. (2023). Confidence intervals for the coefficient alpha difference from two independent samples (groups). *The Quantitative Methods for Psychology*, *19*(2), 194–216. doi: 10.20982/tqmp.19.2.p194.

Received: 31/12/2021 ∼ Accepted: 14/06/2023

The figures follow.

Figure 1 ■ Bonett confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$.

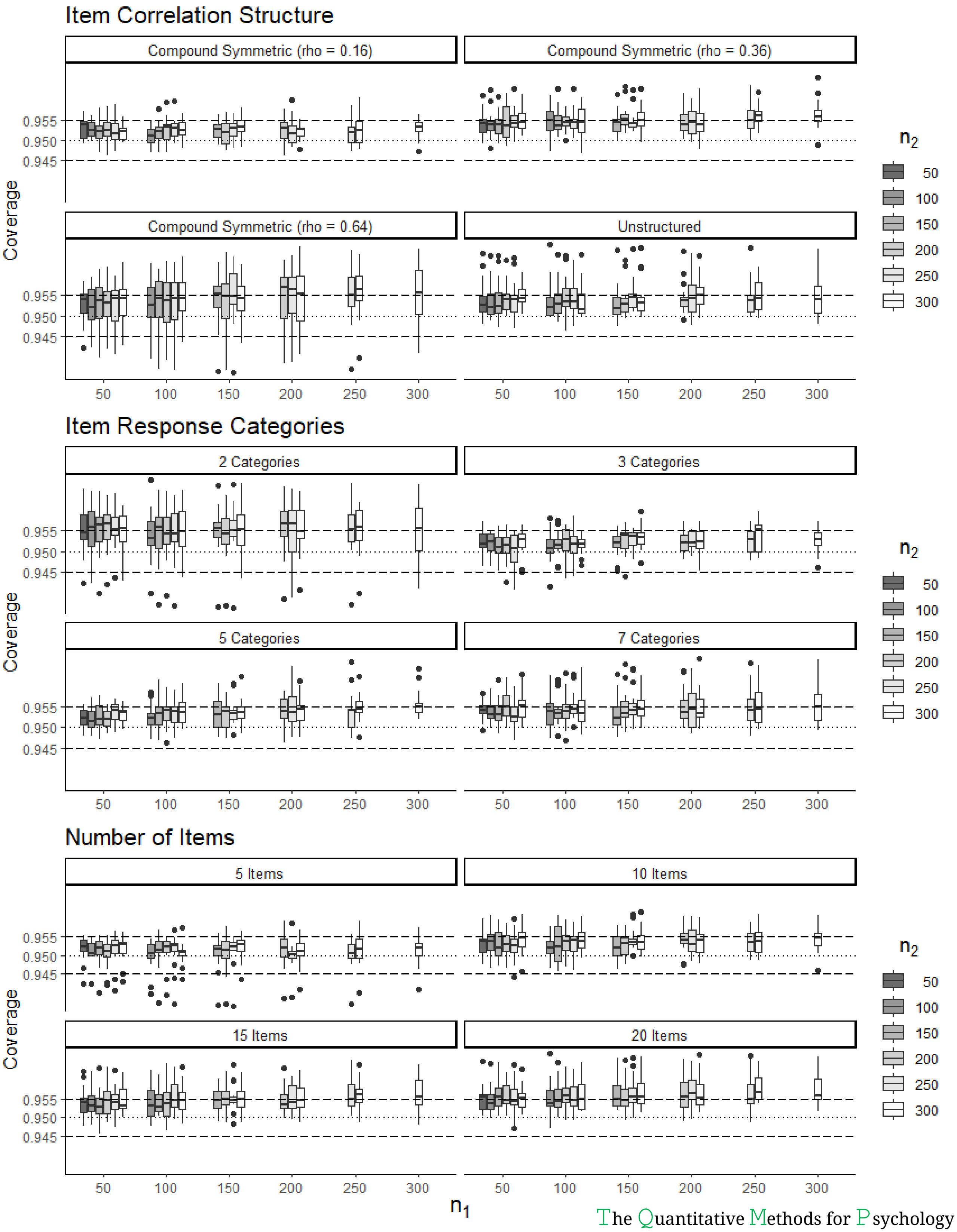


Figure 2 ■ Normal theory 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$.

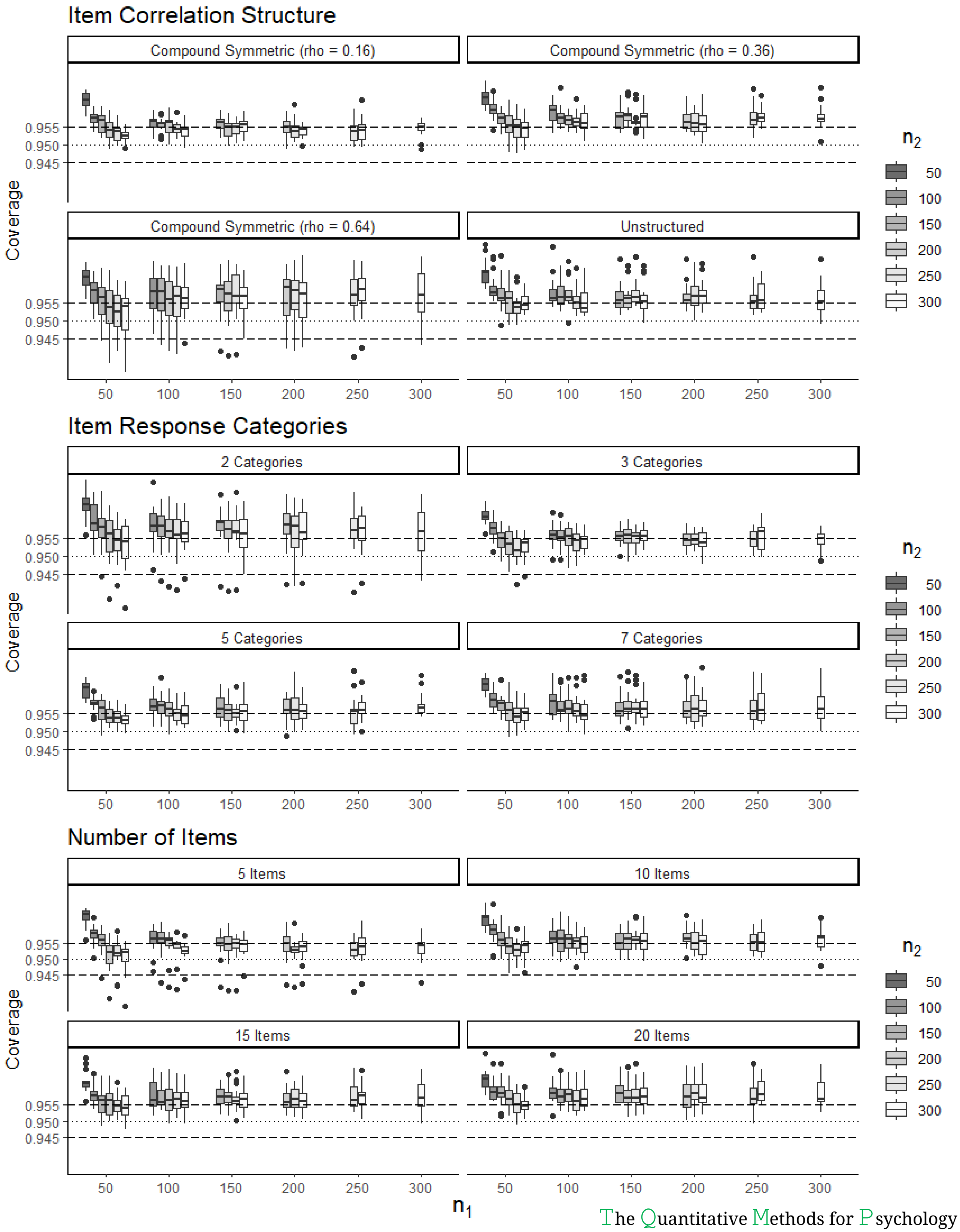


Figure 3 ■ Normal theory bootstrap 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$. n_1 = group 1 & n_2 = group 2 sample sizes; 2,000 bootstrap samples were used.

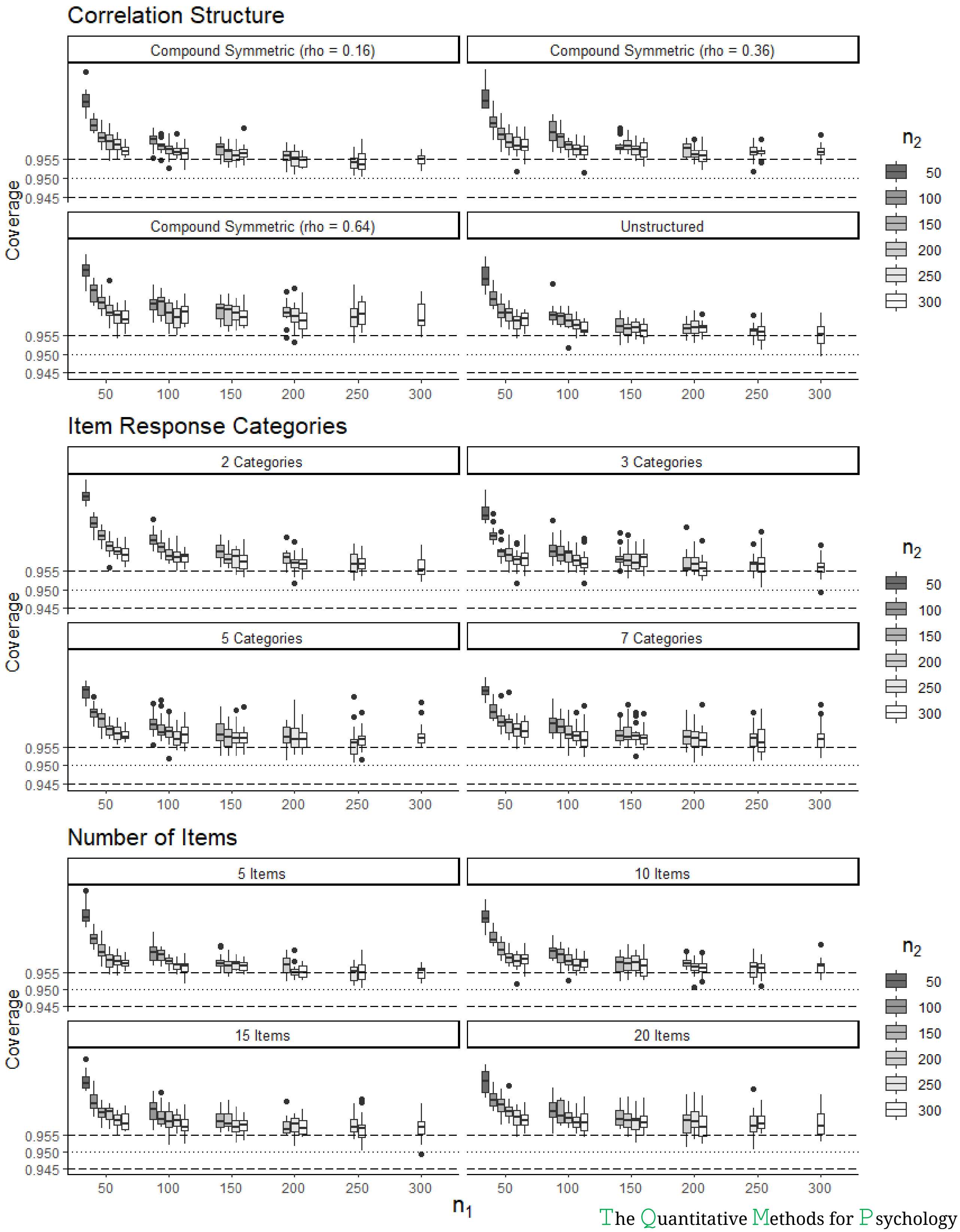


Figure 4 ■ Percentile bootstrap 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$; 2,000 bootstrap samples were used.

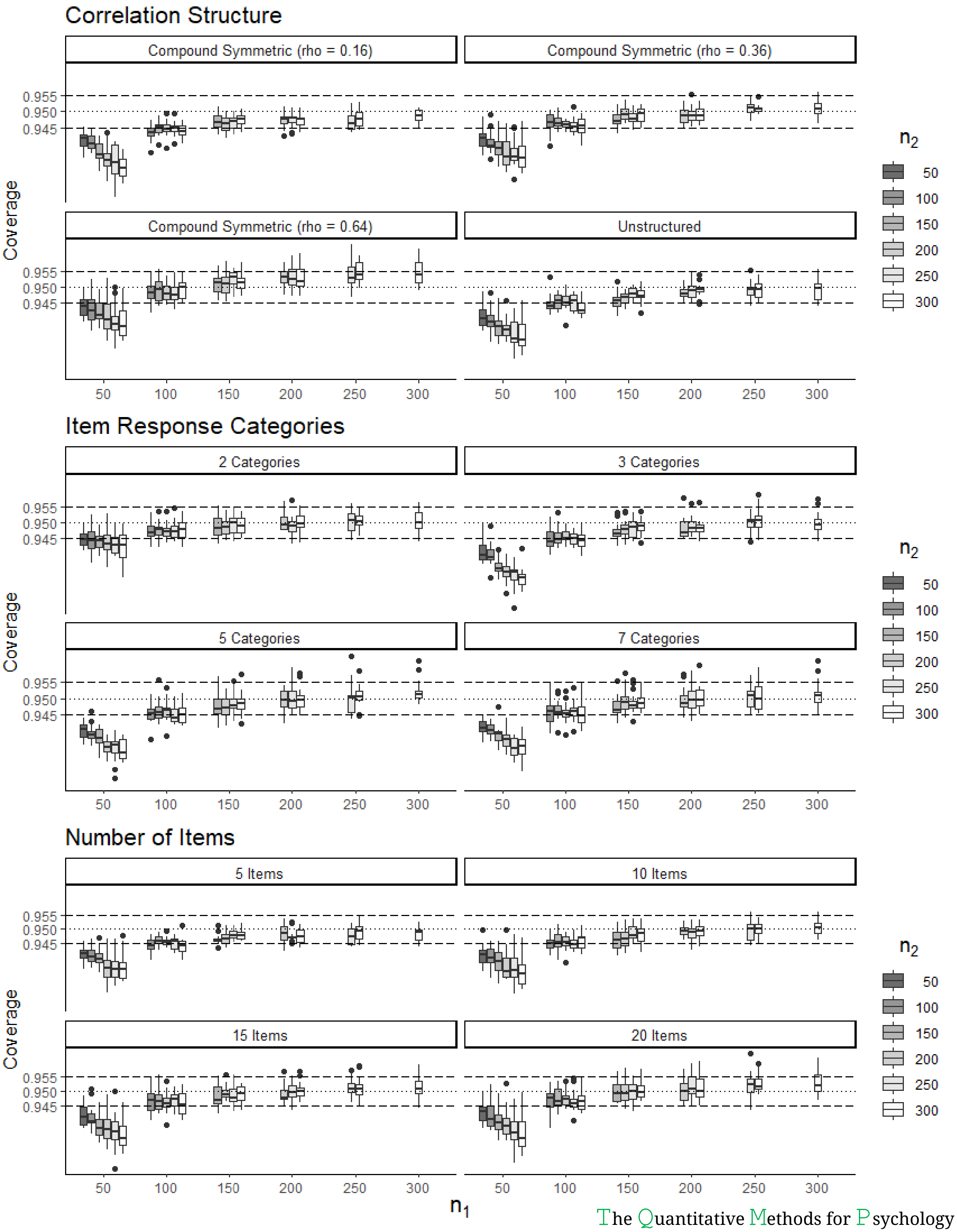


Figure 5 ■ Bias-corrected & accelerated 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.955,\ 0.945]$; 2,000 bootstrap samples were used.

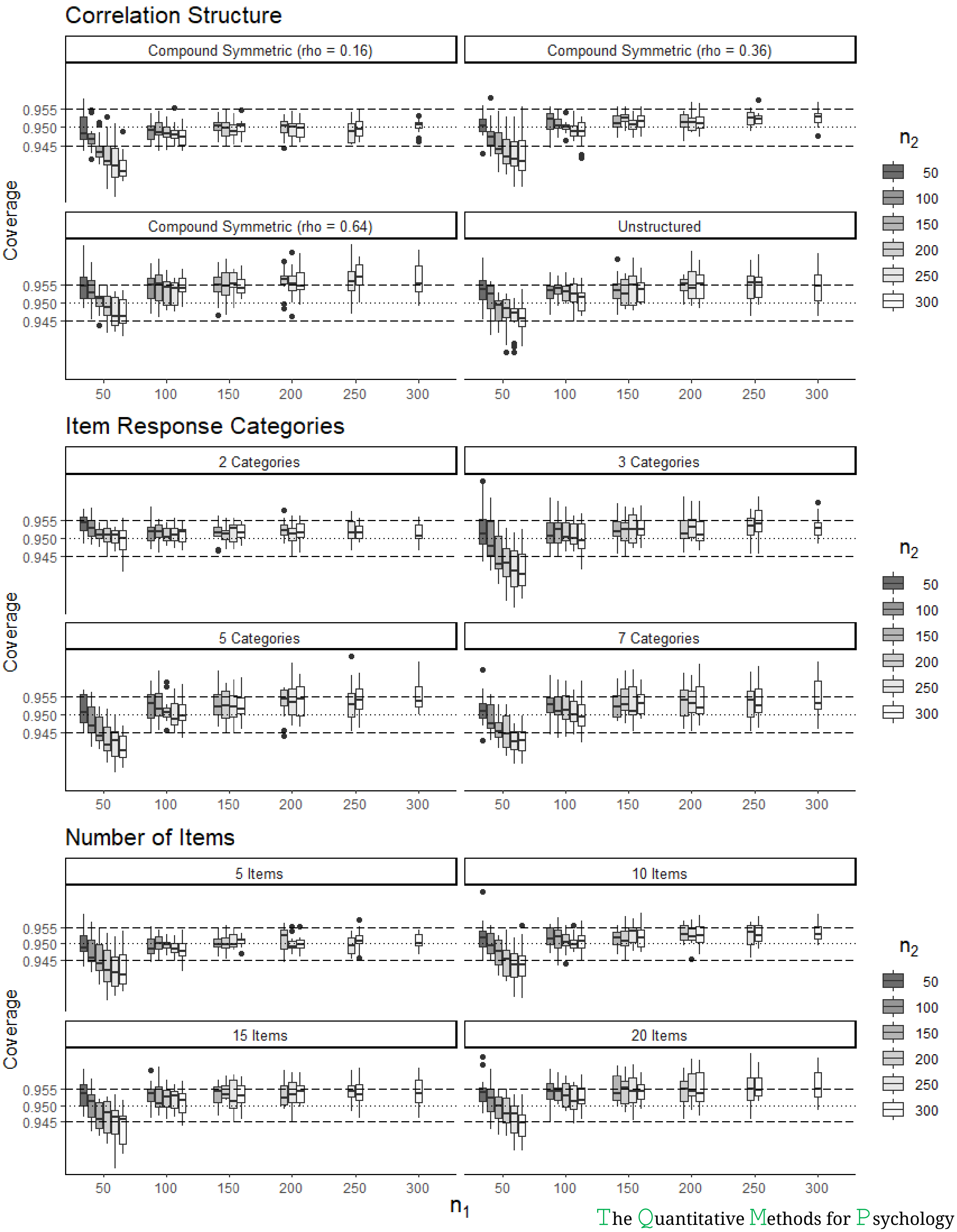


Figure 6 ■ Bootstrap highest density interval 95% coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.955,\ 0.945]$; 2,000 bootstrap samples were used.

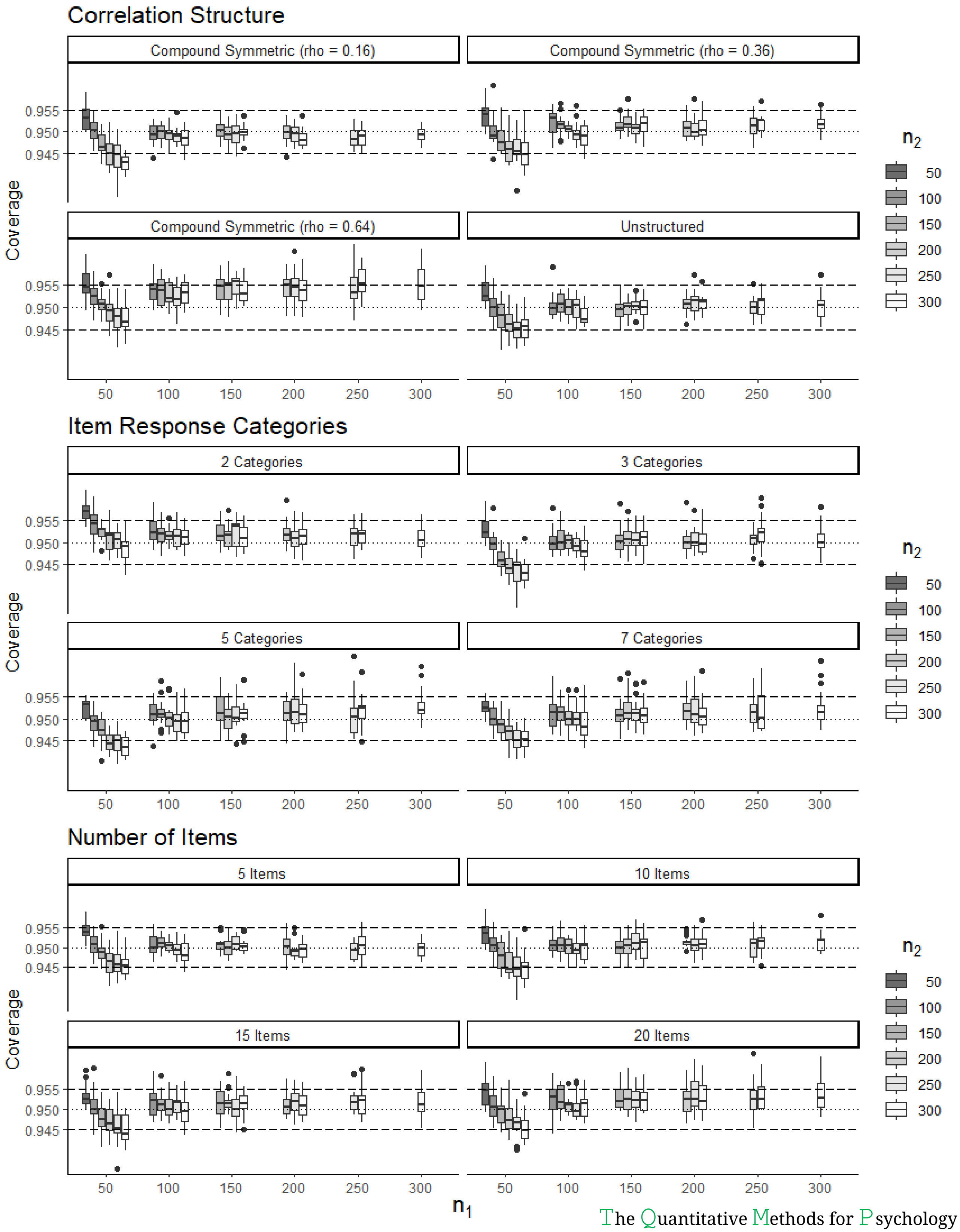


Figure 7 ■ Bonett 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$ with two items. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$.

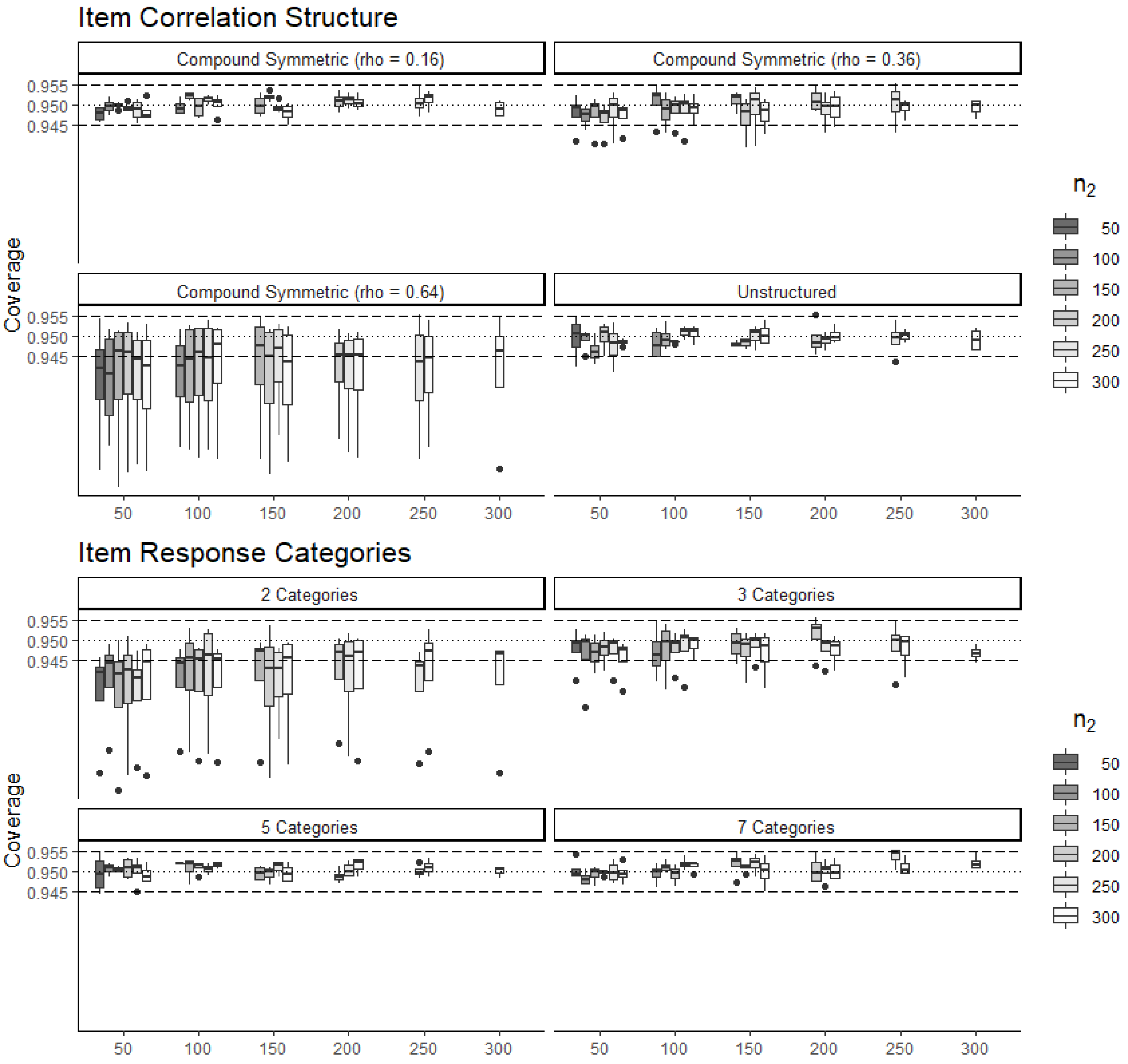


Figure 8 ■ Normal theory 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$ with two items. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$.

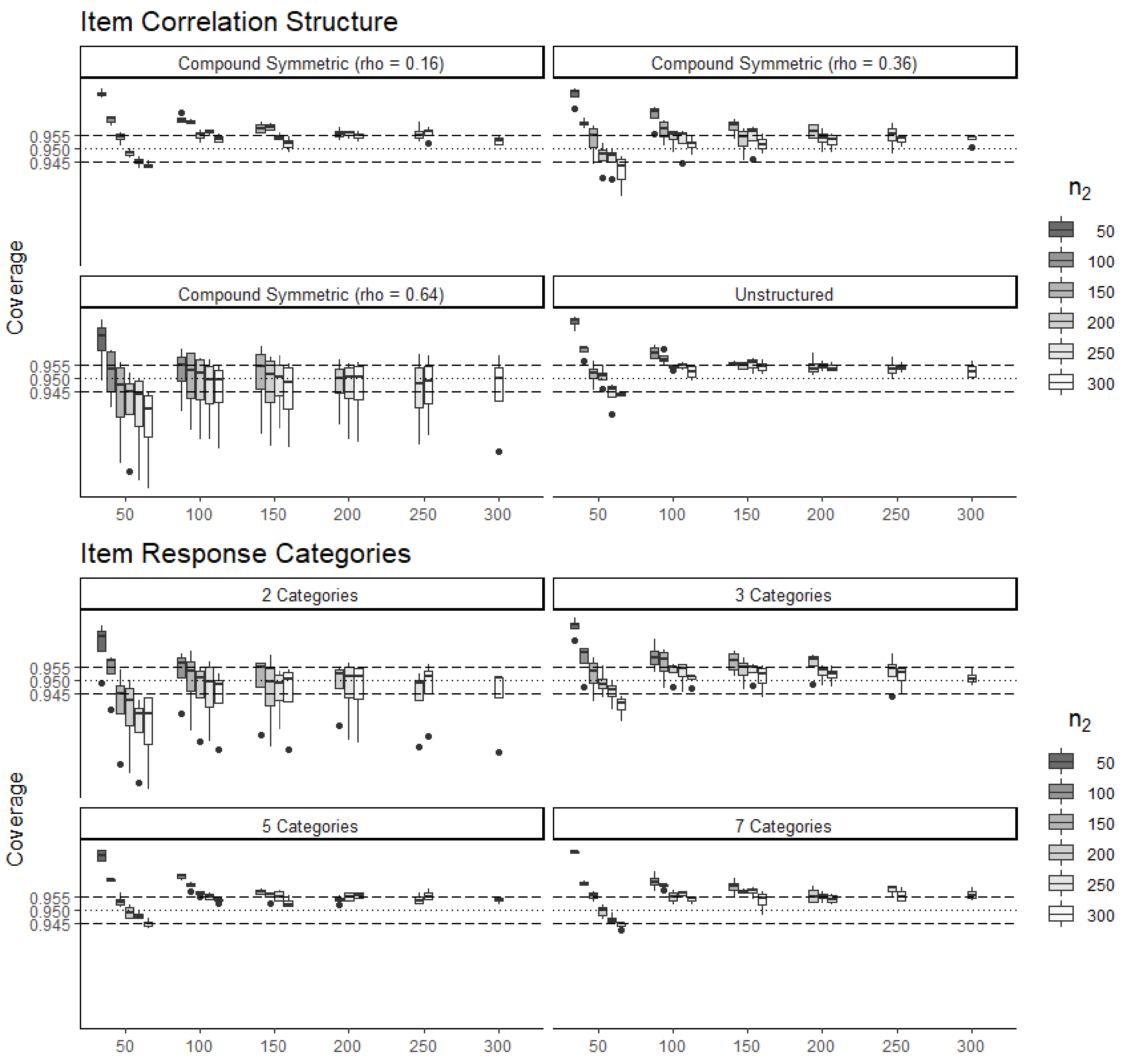


Figure 9 ■ Normal theory bootstrap 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$ with two items. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$; 2,000 bootstrap samples were used.

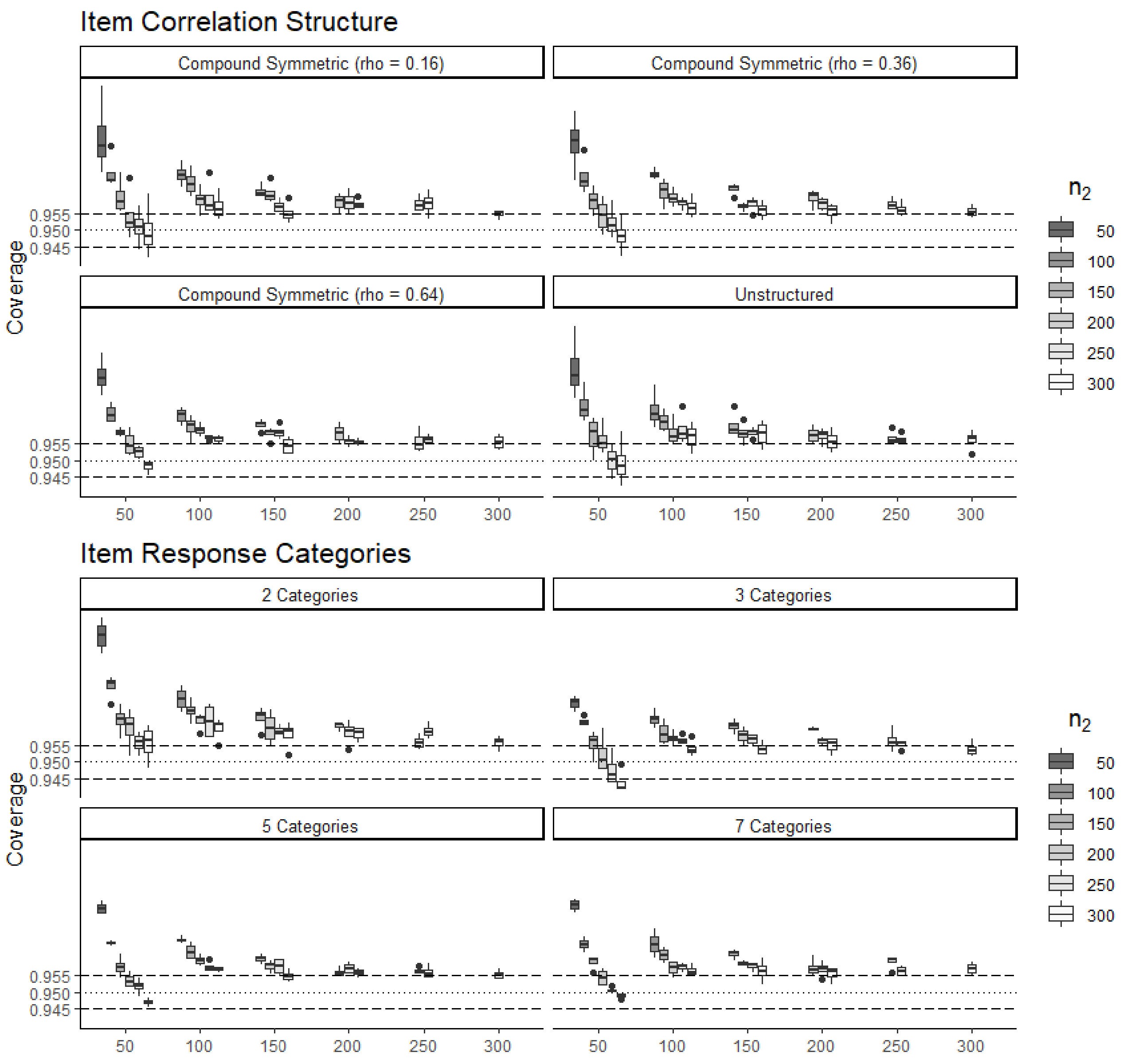


Figure 10 ■ Percentile bootstrap 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$ with two items. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.945,\ 0.955]$; 2,000 bootstrap samples were used.

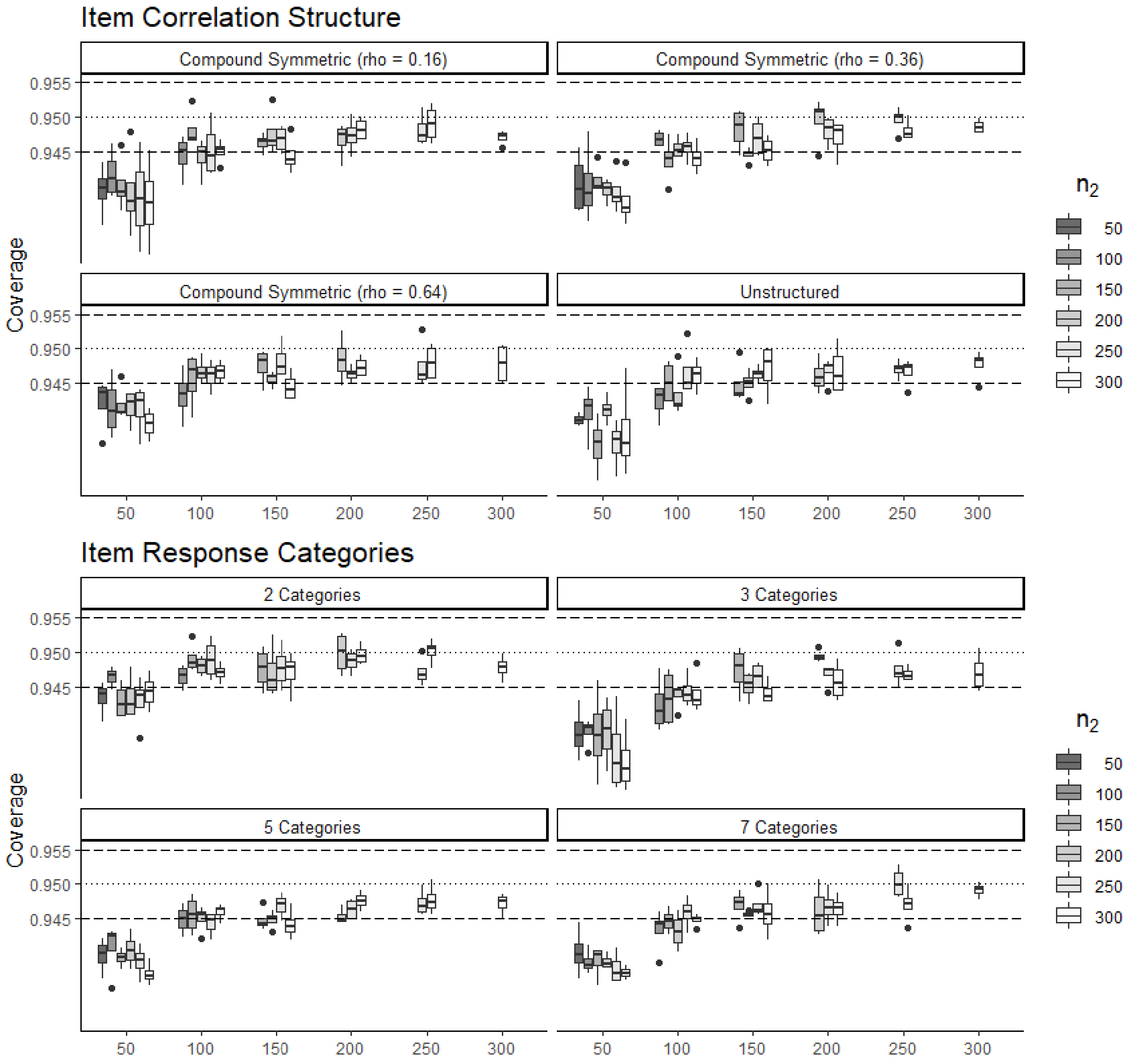


Figure 11 ■ Biased corrected & accelerated 95% confidence interval coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$ with two items. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.955,\ 0.945]$; 2,000 bootstrap samples were used.

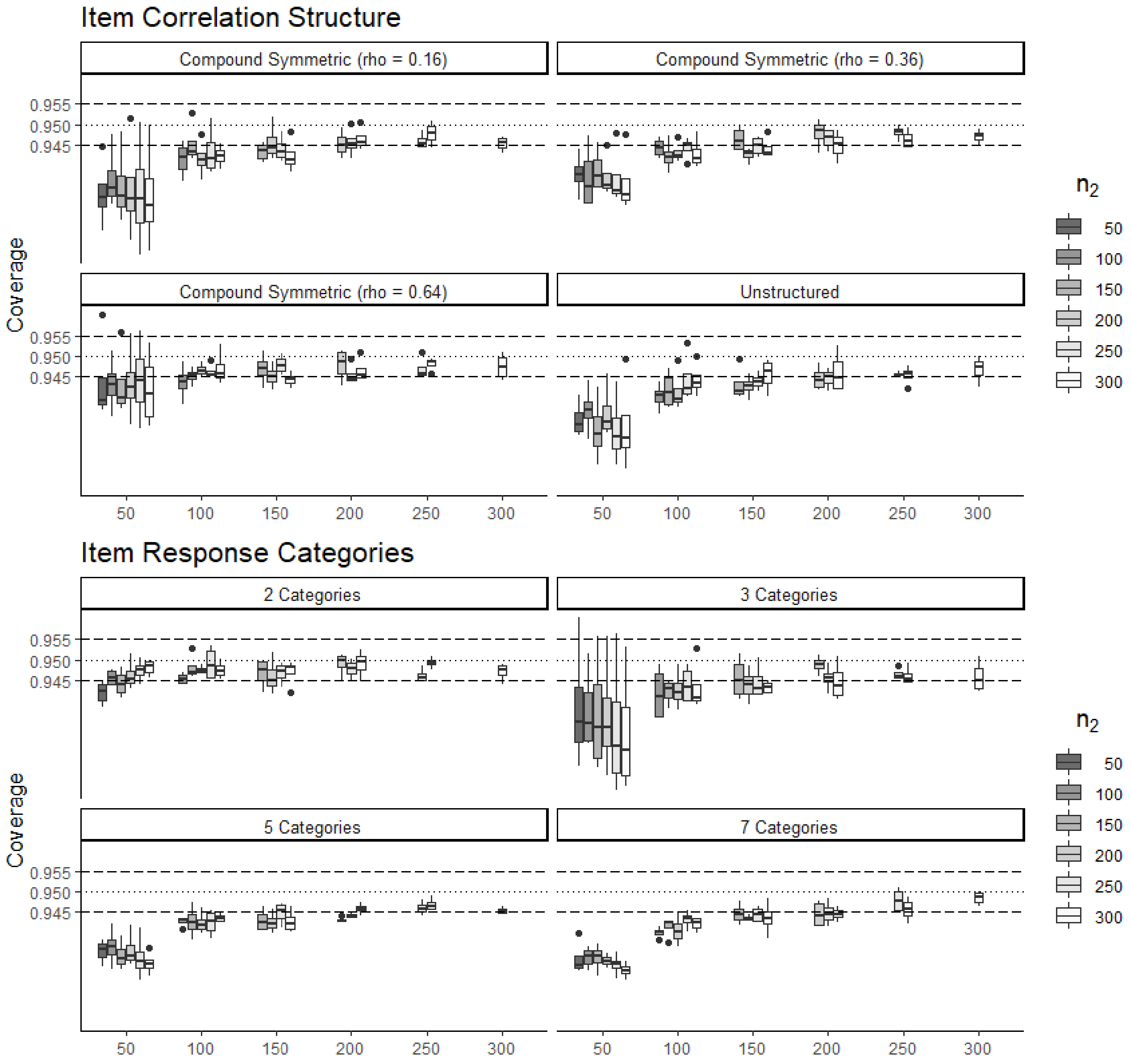


Figure 12 ■ Bootstrapped highest density interval 95% coverage for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2})$ with two items. n_1 = group 1 & n_2 = group 2 sample sizes; acceptable coverage within $[0.955,\ 0.945]$; 2,000 bootstrap samples were used.

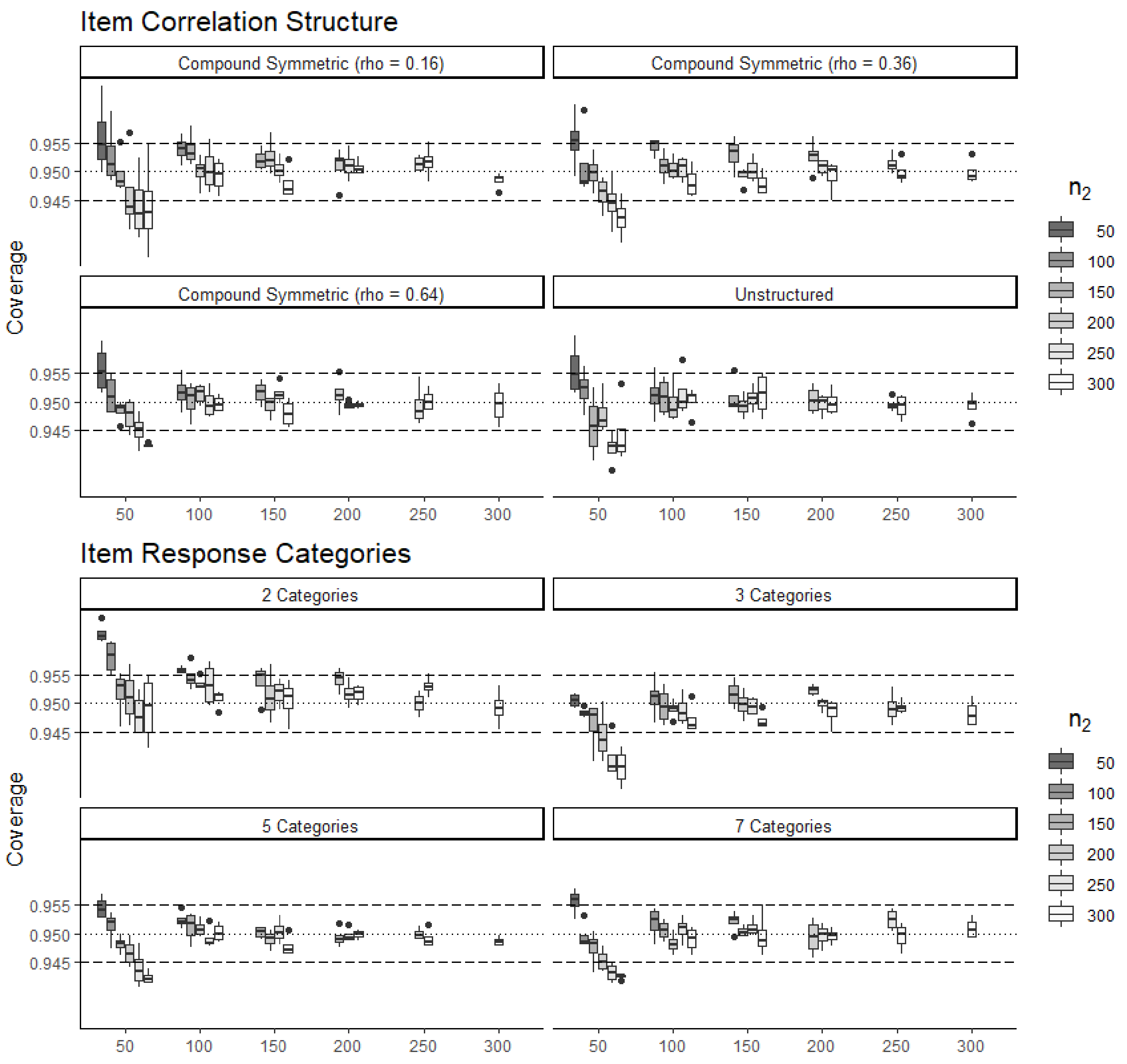


Figure 13 ■ Bias for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2}) - (\alpha_{c1} - \alpha_{c2})$ where n_1 = group 1 & n_2 = group 2 sample sizes.

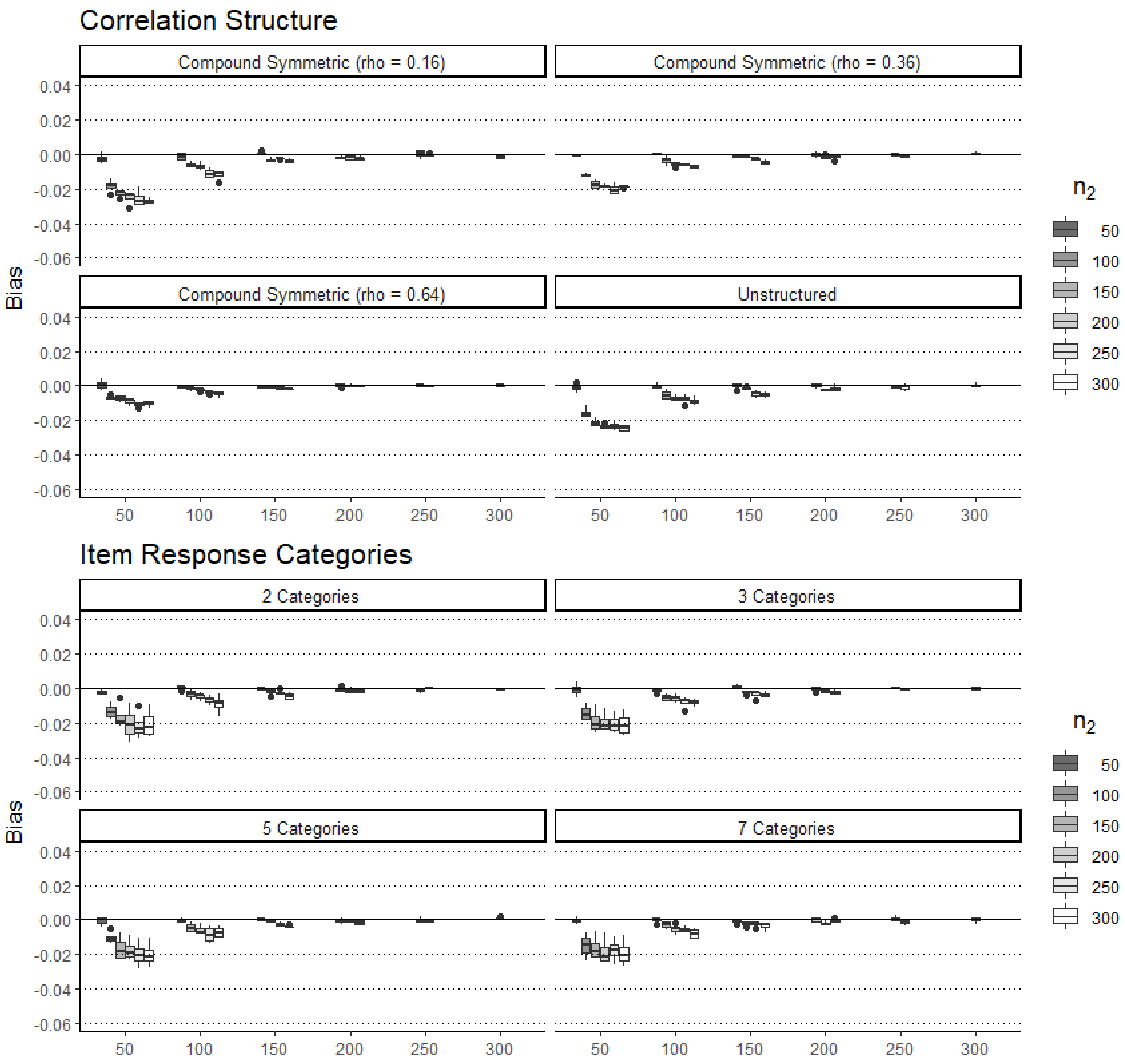
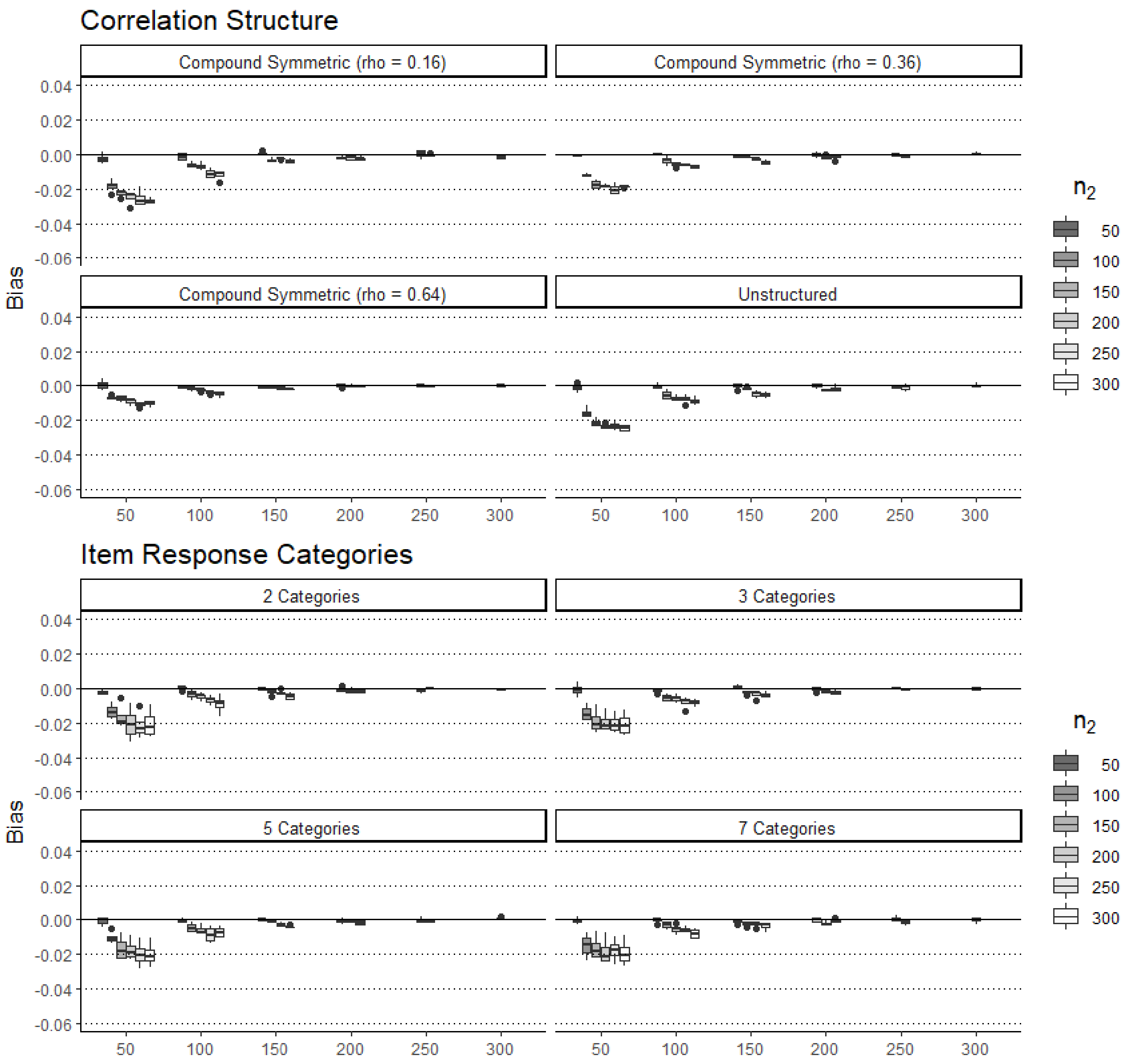


Figure 14 ■ Bias for $(\hat{\alpha}_{c1} - \hat{\alpha}_{c2}) - (\alpha_{c1} - \alpha_{c2})$ with two items where n_1 = group 1 & n_2 = group 2 sample sizes.


Correlation Structure
Compound Symmetric (rho = 0.16)
Compound Symmetric (rho = 0.36)
Compound Symmetric (rho = 0.64)
Unstructured
Bias
n2
Item Response Categories
2 Categories
3 Categories
5 Categories
7 Categories