*Advances in Methodology and Statistics*, 2023, *20*(1), 1–36. https://doi.org/10.51936/vmsz7741

# Latent structure and reliability: A large-scale Monte Carlo study

Josip Novak*, Blaž Rebernjak

*University of Zagreb, Faculty of Humanities and Social Sciences, Department of Psychology*

## Abstract

There are numerous reliability coefficients and α is the most popular. It is known that different coefficients can be appropriate in specific conditions and that α should not be used indiscriminately. However, some coefficients and conditions, particularly regarding the latent structure, lacked attention in previous research. In four Monte Carlo simulations, this study compared α, $\lambda_2$, maximized $\lambda_4$, $\lambda_4$-based on locally optimal splits, $\mu_2$, Gilmer-Feldt, Kaiser-Caffrey α, Heise-Bohrnstedt Ω, Joreskog's ρ, $\omega_{\text{total}}$, algebraic greatest lower bound, greatest lower bound based on minimum rank factor analysis in every condition and also hierarchical ω and asymptotic ω hierarchical in multidimensional conditions. Findings suggest each of these coefficients can be useful at least in some conditions. Most differences in performance were observed in congeneric conditions and conditions with up to moderate loadings. Some coefficients were found to be more useful than previously considered. Results are discussed in the context of existing theory and previous Monte Carlo studies.

*Keywords:* psychometrics, reliability, Monte Carlo simulation

## 1. Introduction

α (Cronbach, 1951; Guttman, 1945) is the most popular reliability coefficient. For brevity, the term "coefficient" represents reliability of test scores in a design that uses a single time point. The term can be considered as equivalent to "reliability coefficient" and "measure of reliability.". In theory, if undimensionality holds and errors are uncorrelated, α represents a lower bound to reliability of a composite. If items are additionally τ-equivalent, α is theoretically equal to reliability (Lord & Norvick, 1968). In practical situations, α is influenced by various factors that make it biased and imprecise. Therefore, it should not be used indiscriminately.

The indiscriminate use of α has been met with criticism, primarily due to its limited usefulness. It mostly underestimates and occasionally overestimates reliability (e.g., Cho, 2021b; Green & Yang, 2009a; Sijtsma, 2009; Sijtsma & Pfadt, 2021). Some researchers suggested α should be abandoned (e.g., McNeish, 2018). Raykov and Marcoulides (2019)

*Corresponding author

*Email addresses:* josip.novak92@gmail.com (Josip Novak), brebernj@ffzg.hr (Blaž Rebernjak)

ORCID iDs: (Josip Novak), (Blaž Rebernjak)

argued that $\alpha$ has practical utility under certain empirical conditions and should be used when justifiable, especially because it can yield similar values as some other coefficients (e.g., Savalei & Reise, 2019). Still, Bentler (2021) suggested $\alpha$ is simply a conservative estimate of a lower bound to reliability. Cho (2022) demonstrated there is no universally appropriate coefficient and that many alternative coefficients have been overlooked.

### *1.1. Latent structure and reliability*

Before estimating reliability, it is recommended to examine the latent structure using confirmatory factor analysis (CFA). The latent structure encompasses measurement model type, dimensionality, and error correlatedness. It is possible to determine if the model type is parallel, $\tau$- equivalent, or congeneric. Parallelism holds if item means, standard deviations, item true scores or factor loadings, errors, and correlations with any other construct are equal. $\tau$-equivalence holds if factor loadings are equal, but error variances differ (Lord & Norvick, 1968). If both factor loadings and errors are heterogeneous, the model is congeneric.

According to latent trait theory, a common assumption is that items are unidimensional.[1] Unidimensionality holds if a single-factor model fits and items are locally independent (McDonald, 1981). Measurement model types have mostly been discussed for conditions in which unidimensionality holds, but there are also multidimensional parallel and multidimensional $\tau$-equivalent models (Cho, 2016). Multidimensional models are increasingly often used (e.g., Gessaroli & Champlain, 2005).

Local dependence refers to the presence of correlated errors or residuals when the underlying latent variable is controlled. Correlated residuals stem from many different sources, such as subgroups of items being associated with different stimulus materials, item arrangement, or transient errors (Green & Yang, 2009a). Sets of locally dependent items make tests multidimensional. However, correlated residuals differ from multidimensionality of purposely measured constructs since the former represents random and the latter represents fixed multidimensionality (Wainer & Thissen, 1996). Correlated residuals may indicate the model has been misspecified (e.g., Shi et al., 2018). If a model requires respecification, residual correlation can be allowed when justifiable (Anderson & Gerbing, 1984).

Reliability coefficients vary in their appropriateness for particular conditions. Violated $\tau$-equivalence and unidimensionality mostly attenuate reliability estimates and correlated residuals can make them biased in either direction (e.g., Raykov, 1998). Coefficients can be broadly divided into those based on item variance-covariance matrix (e.g., Guttman, 1945; ten Berge & Zegers, 1978) and those based on factor analysis (FA; see Bentler, 2021). In theory, coefficients based on item variance-covariance matrix are considered a lower bound to reliability if unidimensionality and uncorrelated errors hold and equal to reliability if also loadings are $\tau$-equivalent (e.g., Guttman, 1945; Novick & Lewis, 1967) with some exceptions (e.g., Gilmer & Feldt, 1983). FA-based coefficients are considered less restrictive as reliability estimators since they incorporate item unique variance and are appropriate for multidimensional structures (e.g., McDonald, 1999).

[1] Unidimensionality refers to a single-factor solution that fits. In the classical test theory (CTT), there is no formal assumption of unidimensionality since each item can contain unique variance, or the assumption of uncorrelated errors since models that assume correlated errors lead in different approaches to reliability (e.g. Sijtsma & Pfadt, 2021).

### 1.2. *An overview of some reliability coefficients and previous research related to the latent structure and reliability*

$\alpha$ was introduced as $\lambda_3$ in a series of six lower bounds to reliability ($\lambda_{1-6}$; see Guttman, 1945). $\lambda_2$ and $\lambda_{4(\max)}$, or maximized split-half (where halves are denoted as *a* and *b*), are considered superior to $\alpha$. The relationship between $\lambda_2$ and $\alpha$ is $\alpha \leq \lambda_2$. The relationship between $\lambda_4$ and $\alpha$ is $\alpha \leq \lambda_{4(\max)}$ and $\alpha = \mathrm{E}[\lambda_{4(\max)}]$. Hunt and Bentler (2015) introduced a $\lambda_4$ variant based on cumulative proportions of locally optimal splits that are below a particular point $Q$ in the distribution of split-halves. Hunt and Bentler (2015) demonstrated their variant is superior to $\alpha$ and $\lambda_{4(\max)}$ under one-factor and two-factor solutions in the case of both parallel and congeneric measurement models with $Q$ points of 0.05, 0.50, and 0.95, especially the $Q$ point of 0.05. In the formulas (1.1), $J$ represents the total number of items, $j$ represents a single item, $X$ represents the total observed score, while Var is variance, which is calculated as the average of squared differences between the mean and each value:

$$\begin{aligned}\lambda_3 &= a = \frac{J}{J-I}\left(1-\frac{\sum \mathrm{Var}_j}{\mathrm{Var}_X}\right)\\ \lambda_2 &= 1-\frac{\sum \mathrm{Var}_j}{\mathrm{Var}_X}+\frac{\sqrt{\frac{J}{J-1}\sum \mathrm{Cov}^2_{j,j-1}}}{\mathrm{Var}_X}\\ \lambda_{4(\max/Q)} &= 2\left(1-\frac{\mathrm{Var}_a+\mathrm{Var}_b}{\mathrm{Var}_X}\right).\end{aligned} \tag{1.1}$$

Correlated residuals can make $\alpha$ either overestimate or underestimate reliability (e.g., Raykov, 1998) and the effect occurs even if not all the residuals are correlated (Zimmerman et al., 1993). Thompson et al. (2010) compared $\alpha$ and $\lambda_{4(\max)}$ under different latent structures, across various sample sizes and test lengths, and concluded $\lambda_{4(\max)}$ can overestimate reliability.

ten Berge and Zegers (1978) presented the $\mu$-series. $\mu_0$ is equivalent to $\alpha$ and $\mu_1$ is equivalent to $\lambda_2$. The coefficients belong to an infinite series of lower bounds to reliability without notable improvement beyond $\mu_2$:

$$\mu_2 = \sum \mathrm{Cov}_{j,j-1} + \frac{\sqrt{\sum \mathrm{Cov}^2_{j,j-1} + \sqrt{\frac{J}{J-1}\sum \mathrm{Cov}^4_{j,j-1}}}}{\mathrm{Var}_X}$$

Gilmer and Feldt (1983) presented a coefficient (GF) for unidimensional congeneric models as a superior alternative to $\alpha$. GF is based on variances and inter-item covariances of $j$ items and $i$ parts, where $l$ represents the row of the inter-item variance-covariance matrix with the largest sum:

$$\mathrm{GF} = \left(\frac{\left(\sum \frac{\mathrm{Cov}_{j,i}-\mathrm{Cov}_{j,l}-\mathrm{Var}_j}{\mathrm{Cov}_{l,j}-\mathrm{Cov}_{i,l}-\mathrm{Var}_l}\right)^2}{\left(\sum \frac{\mathrm{Cov}_{j,i}-\mathrm{Cov}_{j,l}-\mathrm{Var}_j}{\mathrm{Cov}_{l,j}-\mathrm{Cov}_{i,l}-\mathrm{Var}_l}\right)^2-\sum\left(\frac{\mathrm{Cov}_{j,i}-\mathrm{Cov}_{j,l}-\mathrm{Var}_j}{\mathrm{Cov}_{l,j}-\mathrm{Cov}_{i,l}-\mathrm{Var}_l}\right)^2}\right)\left(\frac{\mathrm{Var}_X-\sum \mathrm{Var}_j}{\mathrm{Var}_X}\right).$$

Osburn (2000) compared various coefficients and concluded $\lambda_4$ is typically the least biased estimate of reliability, followed by GF. $\lambda_4$ was also shown to be relatively unbiased in the two-dimensional model, unlike other coefficients. He demonstrated $\alpha$ underestimates reliability under severe violation of $\tau$-equivalence.

Coefficients based on FA are often referred to as ω-family or composite reliability. α is a special case of ω-family if and only if unidimensionality and τ-equivalence hold. ω-family relies on the FA method and is calculated using *j* item loadings ($\lambda$[2]) and FA residuals (θ). Some ω-family coefficients are based on alpha FA and standardized loadings (KC; Kaiser & Caffrey, 1965) or unstandardized loadings (ρ; Jöreskog, 1971). ω based on group factor loadings with CFA is denoted as ω total ($\omega_t$; Revelle & Zinbarg, 2009). However, some coefficients in the ω-family use either group factor loadings, such as $\omega_t$ and principal components-based Ω (HB; Heise & Bohrnstedt, 1970), or general factor loadings, such as hierarchical ω ($\omega_h$; McDonald, 1999) and asymptotic hierarchical ω ($\omega_a$; Revelle, 2022):

$$\omega = \frac{\left(\sum \lambda_j\right)^2}{\left(\sum \lambda_j\right)^2 + \left(\sum \theta_j\right)}. \tag{1.2}$$

α and ω often yield relatively similar values in practical situations if unidimensionality holds, residuals are uncorrelated, and loadings are homogeneous (Savalei & Reise, 2019). The discrepancy between α and reliability is higher when most loadings are low (e.g., Green & Yang, 2009b) or the departure from τ-equivalence is severe, even for a single item (Raykov, 1998).

Jackson and Agunwamba (1977) presented the greatest lower bound (GLB). It corresponds to the smallest average true score variance under the restriction that no individual true score variance is larger than its corresponding observed score variance while keeping the item variance-covariance matrix positive semi-definite over iterations. In theory, it outperforms other existing coefficients and is considered an improvement to the λ-series. It can be calculated using the algebraic approach ($\text{GLB}_\text{A}$; Moltner & Revelle, 2020) or minimum rank factor analysis ($\text{GLB}_\text{M}$; Shapiro & ten Berge, 2002). GLB does not require unidimensionality and τ-equivalence but requires uncorrelated residuals (*e*[3]):

$$\text{GLB} = 1 - \frac{\sum e}{\text{Var}_X}.$$

Although GLB is considered useful with large sample sizes and is occasionally superior to coefficients that require unidimensionality (Sijtsma, 2009), it tends to overestimate reliability (e.g., Hunt & Bentler, 2015) and some authors doubt it is superior to ω (e.g., Revelle & Zinbarg, 2009).

Trizano-Hermosilla and Alvarado (2016) compared α, $\omega_t$, and two GLB algorithms under unidimensional τ-equivalent and congeneric models with uncorrelated residuals. They also varied sample size, test length, and increasing proportion of asymmetrical items. They concluded ω is superior to α in general. Also, GLB is superior to ω with a high proportion of asymmetrical items. Finally, GLB overestimates reliability in some conditions, especially with a small sample.

Edwards et al. (2021) took over conditions from Green and Yang (2009a). They compared α, $\omega_t$, $\omega_{\text{Revelle}}$, $\omega_h$, and GLB under the unidimensional model with uncorrelated residuals and varied sample size, test length, population reliability, and factor loadings. α and ω yielded the least biased estimates of reliability while GLB consistently overestimated reliability.

[2] The symbol λ represents factor loadings in this formula and should be discerned from the λ-series.

[3] While θ represents residuals within the FA framework, *e* is a generic symbol for residuals that encompasses the residuals for both GLB variants.

Trizano-Hermosilla et al. (2021) compared $\alpha$, $\omega_t$, $\omega_h$, $\omega_a$, and two GLB algorithms under bi-factor structures. They varied general factor loadings, specific factor loadings, sample size, and test length. Asymptotic $\omega_a$ followed by $\omega_h$ were shown to be the most accurate for the general factor reliability estimation and outperformed the remaining four coefficients. For total reliability estimation, $\omega_t$ followed by GLB algorithms and $\alpha$ were shown to be the least biased.

Xiao and Hau (2023) compared $\alpha$, ordinal $\alpha$, $\omega_t$, Revelle's $\omega_t$, $\omega_h$, and GLB in unidimensional conditions with uncorrelated residuals. They varied scale (continuous, ordinal), distribution, and factor loadings. Findings indicated that bias was acceptable for continuous scales with varying degrees of non-normality if the loadings were high, but the bias increased with increasing non-normality and moderate factor loadings. For the ordinal scale, most coefficients except $\omega_h$ were acceptable with non-normal data having at least four points.

Cho (2022) reanalyzed several existing Monte Carlo simulation studies to determine why their conclusions differ. He suggested the examination of a small number of coefficients is the main reason for different conclusions in the existing simulation studies.

### *1.3. Current research*

This study extends previous research related to the latent structure in terms of the investigated coefficients and conditions. Moreover, it serves as a follow-up on a recently published paper (see Novak & Rebernjak, 2023), which has a degree of overlap with this study in terms of coefficients. It yielded numerous insights about the differences in coefficient performance and the interactions among the factors related to empirical conditions. However, the primary focus of that paper was limited to factors related to empirical conditions, while this research is focused on the effect of the factors related to the latent structure on reliability estimates and expands the existing knowledge in that regard.

Considering previous research on the effect of latent structure on reliability, it appears that $\omega$-family has been investigated up to a certain point, but the emphasis was predominantly on $\omega_t$ and $\omega_h$. Therefore, KC, $\rho$, and HB will be included in unidimensional conditions, and $\omega_a$ will be included in multidimensional conditions. The study will also examine $\lambda_2$, $\lambda_3/\alpha$, $\lambda_{4(\max)}$, $\lambda_{4(Q=0.05)}$, $\mu_2$, GF, and GLB variants in every condition. Although an analytical approach may not be suitable in this context, the findings are expected to contribute to the existing analytical theory. The results will be compared to previous research and it is expected some previous findings will be replicated, while many insights will be new. Most applicable insights will be distilled into specific recommendations at the end of the paper, providing guidance for practical reliability estimation.

The research is separated into two studies, each consisting of two scenarios. The conditions in the first study represent random multidimensionality in the form of residual correlation and the conditions in the second study represent fixed multidimensionality in the form of correlated factor-structures (Wainer & Thissen, 1996). The first scenario of the first study focuses on unidimensional solutions with $\tau$-equivalent and congeneric models and either uncorrelated or correlated residuals when every residual pair is correlated. Conditions in the second scenario of the first study are specified to cover $\tau$-equivalent models when a proportion of residual pairs are correlated in either direction. The first scenario of the second study primarily focuses on multidimensional conditions involving correlated factors with varying levels of factor correlation and factor loadings. In the second scenario of the second study, the focus is on latent structure with discrete scale and asymmetrical item distribution included to explore their interaction. Therefore, the final scenario involves the number of factors, factor loadings, scale, and item distribution as factors. Every scenario also has sample size and test length as factors because the former affects the precision of every parameter estimation, while the latter is theoretically and practically relevant in the context of reliability estimation. The design for every scenario is displayed in Listing A1.

## 2. Method

Data generation and analysis were conducted using R software (R Core Team, 2022). The description of the data generation procedure is provided alongside each study. The seed was selected using a random number table in each scenario. Mersenne-Twister random number generator was used to generate the data. In the data generation process, packages `randtests` (Caeiro & Mateus, 2022), `Johnson` (Santos Fernandez, 2014), and `whitening` (Strimmer et al., 2022) were used. The whole procedure was parallelized and some functions were compiled to increase the execution speed using base R packages `parallel` and `compiler`. $\alpha$, $\mu_2$, GF, KC, and HB were calculated using package `unirel` (Cho, 2021a). $\lambda_2$, $\lambda_{4(\max)}$, $\lambda_{4(Q=0.05)}$, and $\omega_t$ were calculated using package `Lambda4` (Hunt, 2019). $\text{GLB}_\text{M}$, $\text{GLB}_\text{A}$, $\omega_h$, and $\omega_a$ were calculated using package `psych` (Revelle, 2022). One thousand repetitions were used in every scenario, which was judged as sufficient since the designs are not complex.

In every scenario, performance was evaluated using median bias and bias distribution. The width of the latter is referred to as precision. Median bias was used since some coefficients are expectedly biased in a particular direction, which might result in a non-normal bias distribution. Bias distribution of each coefficient is displayed using box plot. In each figure that displays bias, the dashed vertical line represents the benchmark calculated for each model. Visualization was done using packages `ggplot2` (Wickham, 2016) and `cowplot` (Wilke, 2022).

In line with Cho (2022), some coefficients occasionally yielded values outside the range between zero and unity, mostly in conditions with a small sample. The proportion of such out-of-range values was removed because they are theoretically impossible and would introduce bias into the result presentation. Their frequencies and proportions for each coefficient per scenario are displayed in Table A1.

### 2.1. Study 1

*2.1.1. Conditions.* In the first scenario, conditions are specified to extend Green and Yang (2009a) and Edwards et al. (2021) by comparing additional coefficients and including additional factors. Measurement models for six and 12 items are taken over from Green and Yang (2009a). These measurement models (see Table A2) were selected since they cover various $\tau$-equivalent and congeneric conditions with a range from low to high loadings, some of which can be typically encountered in practice. This study additionally extends these conditions by including correlated residuals. The following levels of residual correlation ($\theta$) are specified: 0 like in Green and Yang (2009a) and Edwards et al. (2021), 0.05 as minor residual correlation that can occur with well-fitting models, and 0.1 as the largest absolute value of residual correlation for close-fitting models (Shi et al., 2018). Including larger residual correlation values was deemed unnecessary because reliability is estimated for the solution that fits and it would introduce additional complexity into the design without adding valuable insights. Residual correlation is identical for every item pair. These residual correlation values also do not result in singular matrices when combined with measurement models from Green and Yang (2009a). Finally, sample size levels are selected to reflect various levels of factor stability: 50, 200, 400 (MacCallum et al., 1999), and 1000. Sample size and residual correlation levels are crossed with every measurement model for each test length condition. Therefore, scenario 1 design is partially crossed and consists of 336 conditions.

In the second scenario, bifactor models for six and 12 items in Green and Yang (2009a) are modified and treated as conditions with pairs of correlated residuals. Therefore, only general factor loadings were treated as contributing to construct-relevant variance, while group factors were treated as correlated residuals. Instead of original general factor loadings 0.3 and 0.8, modified loadings are specified as 0.2, 0.5, and 0.8 for six and 12 items to cover low, medium, and high loadings conditions and make the results more generalizable. Furthermore, instead of residual correlation 0.04 and 0.36, residual correlation ±0.05 and ±0.15 are selected to examine the influence of both positively and negatively correlated residuals. In every measurement model and test length condition, either one or two residual pairs are correlated, which corresponds to certain empirical conditions in which it was justifiable to allow residual correlation (e.g., Anderson & Gerbing, 1984). Sample size levels are identical to the first scenario. Sample size and residual correlation levels are crossed with every measurement model for each test length condition. The design in scenario 2 is fully crossed and consists of 192 conditions.

*2.1.2. Data generation.* The data generation procedure was identical for both Study 1 scenarios. Data were generated by sampling from a continuous normal distribution with $M = 0$ and $SD = 1$. Random normal variables were generated with the following validity check procedure. If the generated data did not meet the target criteria, they were generated anew or transformed. The first part of the validity check involved outlier detection with the conventional threshold of $|z| \geq 3$. Outlier replacement with a random value from normal distribution was done until no outliers were present to control the effect of outlier presence on reliability. Afterward, one-sample Wald-Wolfowitz test was used to test the randomness of the generated data with the 0.05 $p$-value threshold. If the data were non-random, the procedure was repeated until randomness was achieved. Finally, if sample skewness $> 1$, Johnson transformation was applied. Skewness threshold of 1 was selected to control for capitalization on chance in conditions with a small sample. The described procedure was applied to the generation of both true scores and residuals. True scores and residuals were generated separately for each sample size ($N$). True score inter-item correlation for each measurement model was induced by multiplying randomly generated data with Cholesky decomposition of factor loading products matrix. Residuals in the uncorrelated residuals conditions were additionally decorrelated using zero-phase component analysis to remove potential spurious correlation. Residual correlation ($\theta$) was induced by multiplying randomly generated data with Cholesky decomposition of a target correlation matrix. True scores and residuals were summed and $J$ residuals were multiplied with $1 - \lambda_j$ to make the observed scores in line with the FA model. The validity of data obtained using this procedure was checked using FA several times before the data generation procedure to make sure the method produces factor structures in line with previously tested algorithms. Afterward, coefficients were calculated and compared to the benchmark. The benchmark for measurement model ($\lambda$[4]) and test length ($J$) conditions was based on population models as in Green and Yang (2009a). These benchmark values represent population reliability that is typically estimated in practical research. The benchmark is specified according to Equation (1.2), but based on population model loadings, while ω-family coefficients are based on sample loadings. The benchmark represents the CTT view of true reliability that assumes uncorrelated errors, while some conditions in the design represent situations in

[4] $\lambda$ that represents loadings in various measurement models should be discerned from $\lambda$ coefficients.

which sample residuals are correlated. The documented code for the described procedure is provided in the Supplement (online).

*2.1.3. Study 1 results.* Figure 1 and Figure 2 represent coefficient bias in conditions with uncorrelated residuals for combinations of $\lambda$ and $N$ levels. Figure 1 displays $\tau$-equivalent and congeneric factor loading conditions, while Figure 2 contrasts congeneric conditions with an increasing violation of $\tau$-equivalence and congeneric conditions with heterogeneous loadings. Results with uncorrelated residuals are comparable to Green and Yang (2009a) and Edwards et al. (2021).

Figure 1 clearly illustrates coefficients differ in performance even if $\tau$-equivalence holds unless loadings are high ($\lambda = 0.8$), in which case estimates are more precise and less expectedly biased. Moreover, if $\tau$-equivalence holds and loadings are high, coefficients perform quite similarly. If $\tau$-equivalence holds and loadings are not high, coefficients differ in performance and mostly tend to underestimate reliability. If $\tau$-equivalence holds, loadings are low ($\lambda = 0.2$), and $N$ is small, coefficients generally tend to overestimate reliability. In conditions with heterogeneous loadings, estimates do not converge with an increase in $N$ and $J$, in contrast to conditions in which $\tau$-equivalence holds. Expectedly, an increase in $N$ results in improved precision, especially if combined with higher loadings and increased $J$. Finally, an increase in $J$ results in somewhat improved precision in general, even if $N$ is small.

$\alpha$ performs similarly to many other coefficients when $\tau$-equivalence holds, but underestimates reliability the most in conditions with heterogeneous loadings. GLB variants tend to overestimate reliability in some conditions. $\lambda_2$ and $\lambda_{4(\max)}$ did not outperform other coefficients in any condition. When loadings are moderate and $N$ is at least 200, HB and $\text{GLB}_\text{A}$ generally outperform others. However, if $N$ is at least 400, $\text{GLB}_\text{M}$ becomes useful. In the case of low loadings and $J$ being 6, $\rho$ outperforms others up to an $N$ value of 200. On the other hand, GF, $\omega_t$, and HB outperform other coefficients if $N$ is at least 400. KC appears to outperform other coefficients in heterogeneous loading conditions, followed by HB.

Figure 2 illustrates most coefficients tend to underestimate reliability expectedly up to 0.10, which is similar to findings in Figure 1. $\alpha$, followed by $\lambda_2$, underestimates reliability the most in nearly every condition. Only if loadings are all relatively high (0.8) except for one item, $\alpha$ even outperforms some coefficients in terms of median bias, but mostly not in terms of precision if the sample is not large. On the other hand, $\text{GLB}_\text{M}$ tends to overestimate reliability in general. Moreover, coefficients are generally imprecise if the $N$ is small, but their precision increases with an increase in $N$ and loadings, except for GF and $\text{GLB}_\text{M}$ in some conditions. If most loadings are low, $\mu_2$ and $\lambda_{4(Q=0.05)}$ are the least biased. $\omega_t$, HB, KC, and $\text{GLB}_\text{A}$ seem to be the closest estimates of reliability most often, but the latter is useful only if loadings are relatively high and $N$ is not small. KC appears to be the most precise in most conditions. $\lambda_{4(\max)}$ appears to be generally superior to $\lambda_{4(Q=0.05)}$ in terms of median bias but inferior in terms of precision.

Results for 12-item conditions are displayed in Figure A1 due to numerous similarities with six-item conditions. They suggest coefficient precision improves with an increase in $N$ and when these results are compared to Figure 2, it appears an increase in $J$ facilitates the precision. Compared to heterogeneous conditions in Figure 1, it seems coefficients can approximately converge with an increase in $N$ and $J$ even if $\tau$-equivalence is violated, but most loadings have to be high. Estimates are densely distributed around the benchmark only if both $N$ is large and loadings are homogeneous and high. In such situations, there is little difference in coefficient performance.

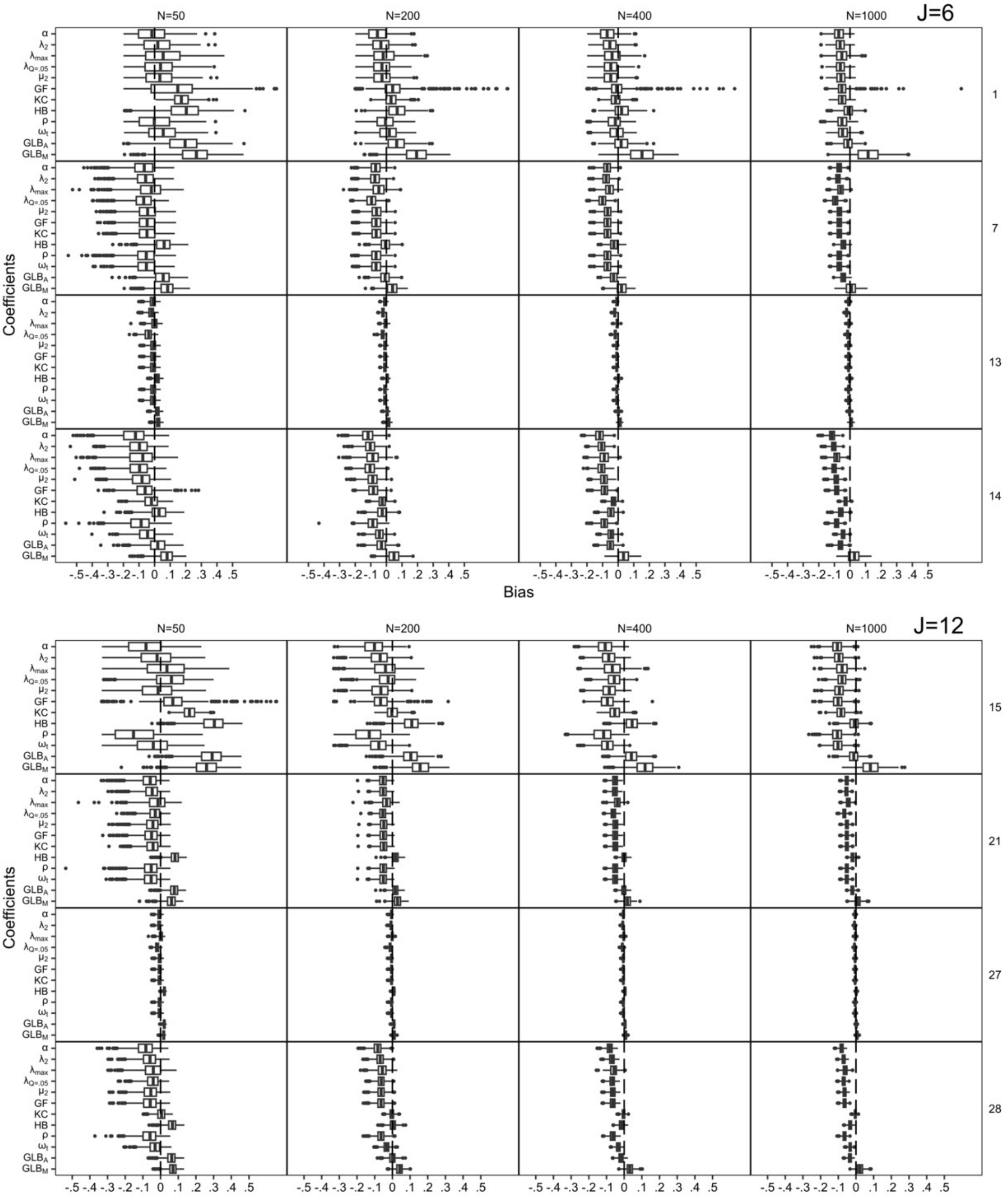

**Figure 1.** Three τ-equivalent conditions and congeneric condition with heterogeneous loadings.

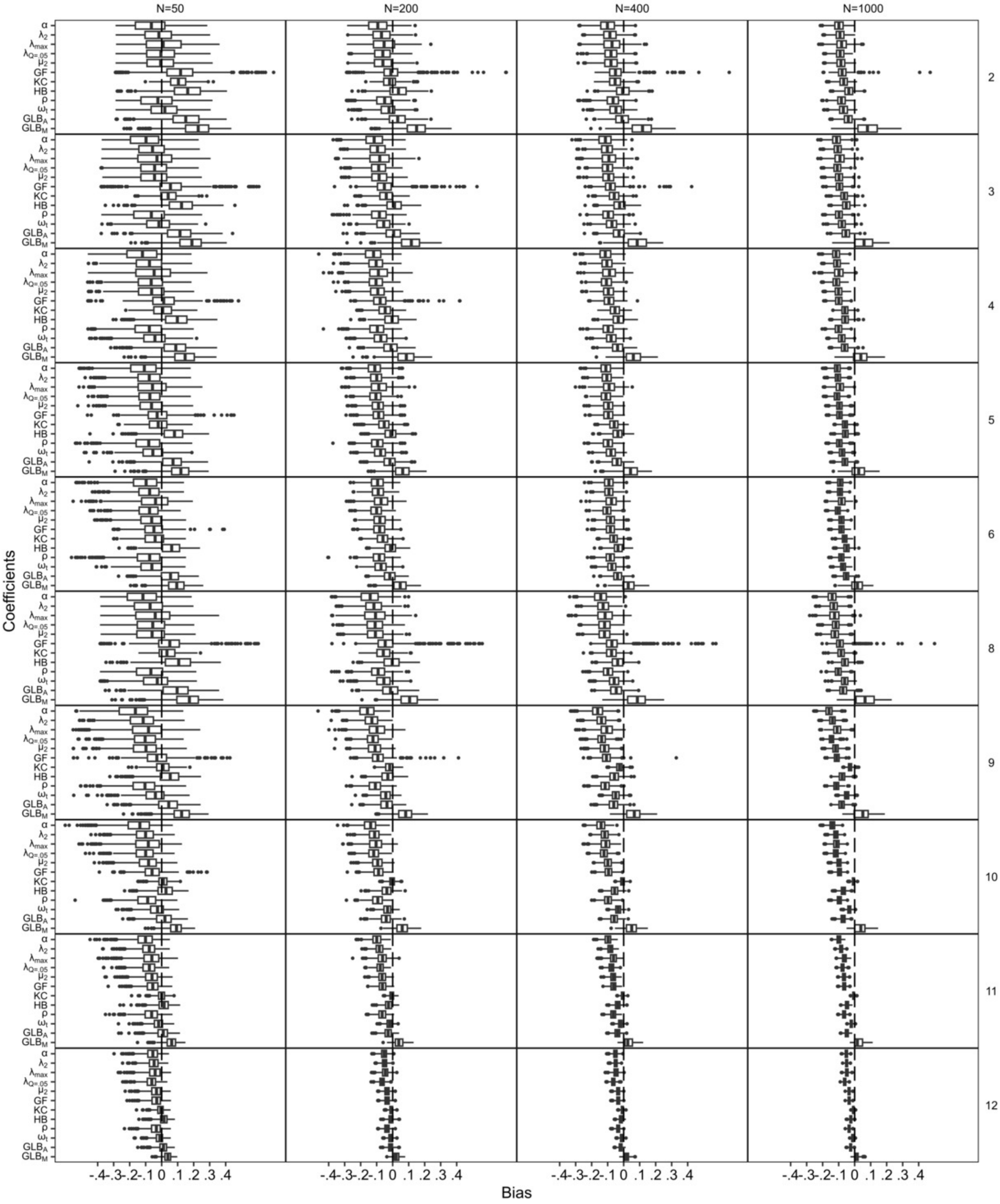

**Figure 2.** Six-item conditions with an increasing violation of τ-equivalence.

There are some specific differences compared to six-item conditions. $\alpha$ also performs the worst in general but performs more similarly to most other coefficients with an increase in $N$. GF can expectedly both underestimate and overestimate reliability and only appears to be useful if most loadings are low and $N$ is small. Unlike in conditions displayed in Figure 2, $\lambda_2$ is not the second worst-performing coefficient in most conditions, but it also never outperforms other coefficients, and neither do $\mu_2$ and $\rho$. $\lambda_{4(\max)}$ appears to be generally superior to $\lambda_{4(Q=0.05)}$ in terms of median bias but inferior in terms of precision unless loadings are up to moderate. KC outperforms other coefficients if there are approximately half low and half moderate loadings, or approximately half low and half high loadings and the sample is 200 to 400. HB performs similarly to KC in terms of median bias but is less precise overall and more biased in more conditions. KC, HB, and $\omega_t$ outperform other coefficients if most loadings are high, KC somewhat more often than $\omega_t$ and HB. HB appears to be the least biased if $N$ is 200 and at least a third of loadings are high, which was not the case in six-item conditions. However, $\lambda_4$ variants outperform KC and $\omega_t$ if loadings are up to moderate and heterogeneous and $N$ is small. GLB variants are expectedly positively biased in most conditions, but $\text{GLB}_\text{A}$ is the least biased if $N$ is 200 and two thirds of loadings are high.

When Figures 1, 2 and A1 are compared, it appears coefficient precision improves more easily with an increase in $J$ if $\tau$-equivalence holds. Otherwise, the improvement in precision is limited even if only one item violates $\tau$-equivalence. However, an increase in $J$ somewhat facilitates improvement in precision regardless of the measurement model, but more easily if most loadings are high.

Figure 3 and Figure 4 represent coefficient bias in conditions where results are averaged over measurement models. This approach aligns to the logic of the Latin hypercube sampling and these results are expected in a range of low to high loadings for both $\tau$-equivalent and congeneric measurement models displayed in Green and Yang (2009a).

Figure 3 demonstrates $\alpha$, $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, and $\omega_t$ tend to underestimate reliability when averaged over measurement models, even if all the residuals are correlated up to 0.10. It appears low average positive residual correlation can occur simultaneously with expected underestimation. KC, HB, and GLBA perform similarly and outperform other coefficients in most conditions, especially the former. However, $\text{GLB}_\text{A}$ requires $N$ at least 200, like in conditions with uncorrelated residuals. $\omega_t$ performs similarly to KC and HB if the residuals are correlated. $\text{GLB}_\text{M}$ is expectedly positively biased in general. Finally, GF is expectedly the most imprecise in every condition, followed by $\text{GLB}_\text{M}$.

Figure 4 illustrates that, compared to six-item conditions in Figure 3, bias is systematically reduced for most coefficients, even if residuals are correlated. GLB variants display a higher tendency to overestimate reliability compared to six-item conditions in Figure 3. KC and HB are among the least biased coefficients, like in Figure 4, but the latter requires $N$ at least 200. However, $\text{GLB}_\text{A}$ is useful in some conditions with correlated residuals if $N$ is at least 400. Also, $\text{GLB}_\text{A}$ requires a larger $N$ if $J$ is higher. $\alpha$, $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, and $\omega_t$ display a similar performance pattern as in six-item conditions.

Finally, Figure 3 and Figure 4 demonstrate large $N$ results in negatively skewed bias distribution if residuals are correlated, and in such conditions, coefficients still expectedly underestimate reliability regardless of $J$, except for $\text{GLB}_\text{M}$. However, it is not certain whether this would be so if the average residual correlation exceeds 0.10. Still, it appears even if the model is borderline misspecified, as indicated by residual correlation, reliability coefficients can underestimate reliability or be unbiased.

It appears at least some factors in the design interact so potential interactions are probed

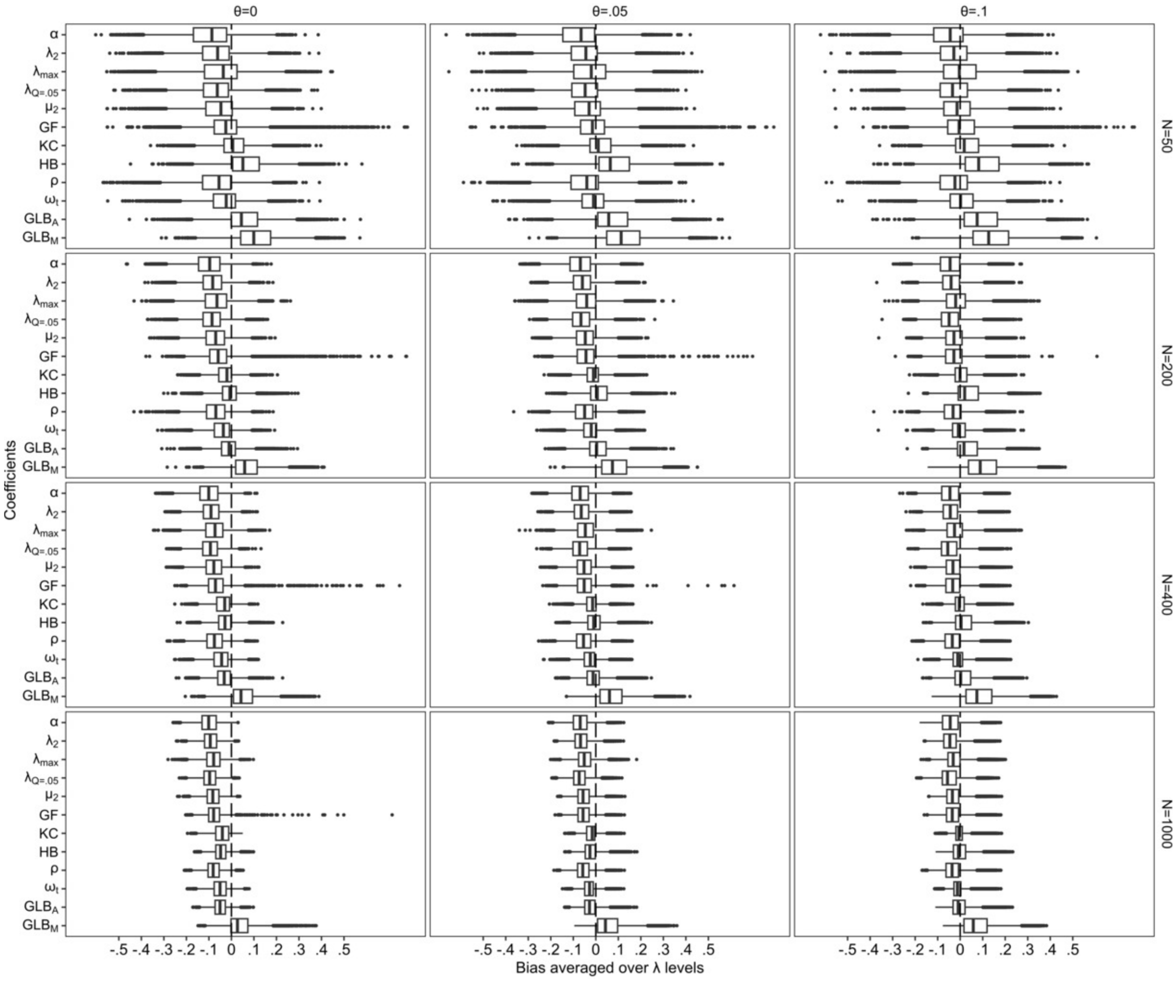

**Figure 3.** Six-item conditions with correlated residuals averaged over measurement models.

using ANOVAs with coefficient bias as the dependent variable and partial $\eta^2$ as the effect size. Every level of each factor was included, except for loadings (λ), in which case only τ-equivalent conditions were retained as factor levels. Conventional thresholds for small (0.01), medium (0.06), and large (0.14) partial $\eta^2$ were used in the interpretation. Effect sizes were considered trivial if below the threshold for medium effect size and only effect sizes that are nontrivial for at least one coefficient are displayed. Main effects and interactions are ordered from left to right based on the global effect size sum to make them ordered by relative importance in influence on reliability estimates, which may not be visible in the display. Effect sizes for each factor and interaction are displayed in Figure A2.

The main insights regarding each factor's impact on bias in reliability coefficients are the following. λ emerges as the most influential factor with a large effect size. θ also exhibits a non-trivial effect, but is somewhat less influential compared to λ. The interaction λ × θ proves to be relevant among the interaction effects. In contrast, the main effects of *N* are either trivial or large. It appears that higher-order interaction effects have limited relevance in this specific combination of factors. Therefore, it appears that λ and *N* are of primary importance for the coefficient selection. However, the specifics of each coefficient should also be considered. For instance, GLB variants, HB, and KC are affected by λ the most, while ρ is affected the least, followed by α.

In scenario 2, it was demonstrated how residual correlation influences reliability estimation if not all the residuals are correlated and if residuals are positively or negatively

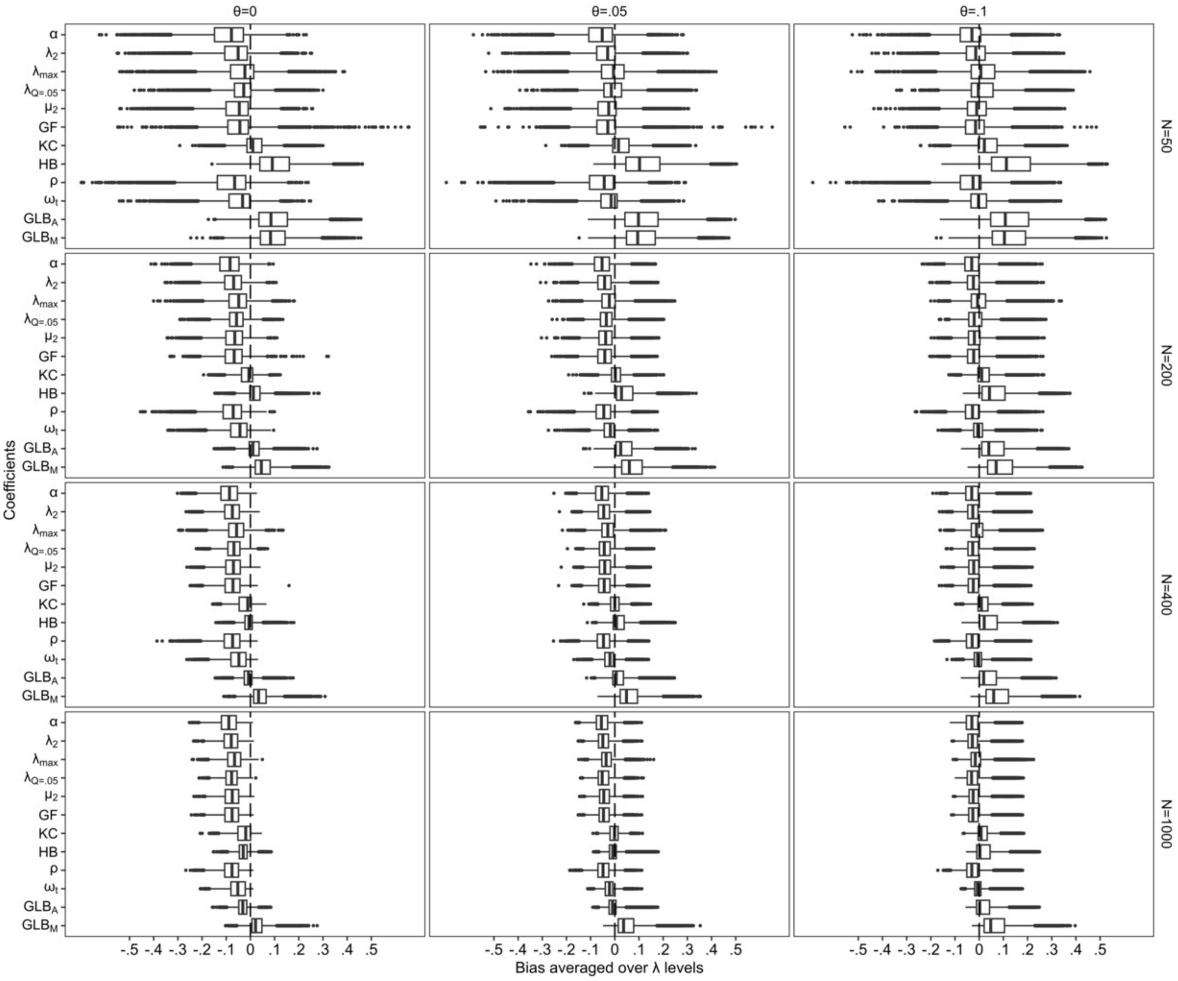

**Figure 4.** 12-item conditions with correlated residuals averaged over measurement models.

correlated. It was observed that residual correlation 0.05 trivially affects reliability estimation, so the display is limited to residual correlation 0.15. Also, since *N* levels merely show different degrees of coefficient stability in this case, the display is limited to sizes 50 and 1000. Results are displayed in Figure 5.

Figure 5 illustrates coefficient precision is improved if λ is 0.8, even if *N* is small. In such cases, it makes little difference which coefficient is used. This is similar to findings for conditions with uncorrelated residuals in Figures 1, 2 and A1, which suggests bias induced by the correlation of residual pairs can be nearly completely mitigated with an increase in λ and *N*. Expectedly, bias can be increased in either direction if two residual pairs are correlated compared to one pair, which is similar for every coefficient and depends on the direction of residual correlation. This increase in bias is larger if *J* is 6. However, most coefficients still expectedly underestimate reliability. Furthermore, pairs of correlated residuals result in peculiar coefficient performance in some conditions. More specifically, if λ is 0.2 and *N* is small, it appears up to two pairs of positively correlated residuals make coefficients underestimate reliability and vice versa. If loadings are small, $\lambda_{4(\max)}$ is the least biased if also *N* is small, while $\text{GLB}_\text{A}$ is the least biased in 12-item conditions if also *N* is large. If loadings are moderate, $\lambda_{4(\max)}$ is the least biased if also *N* is small, while $\text{GLB}_\text{M}$ is the least biased if also *N* is large. It appears these coefficients are useful in conditions when only a proportion of residuals are correlated.

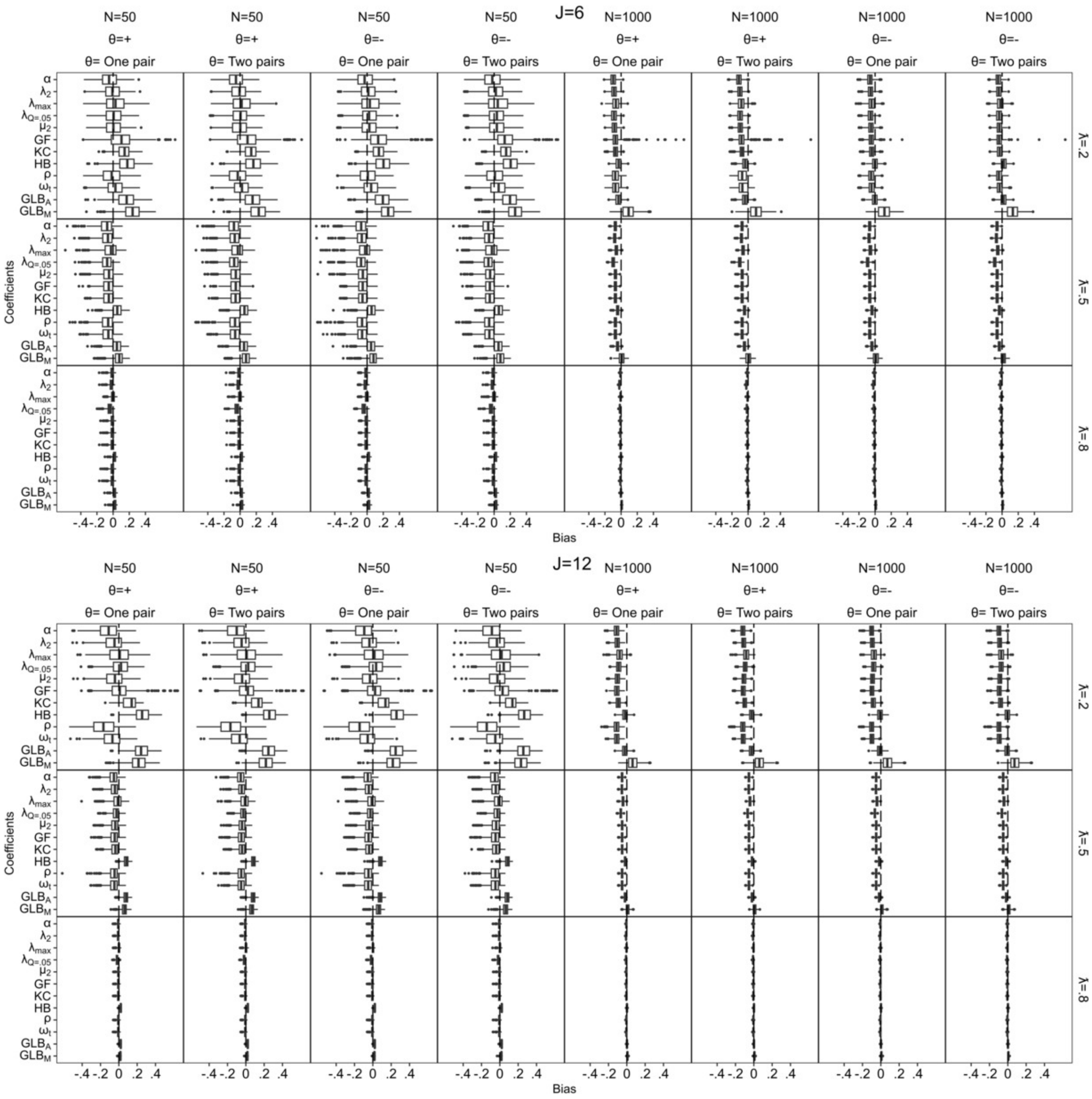

**Figure 5.** Conditions with residual pair correlation 0.15.

*2.1.4. Study 1 findings in the context of previous research.* Study 1 findings are contextualized in Listing 1, which summarizes insights provided in Figures 1–5. The insights are compared to previous findings, starting from Green and Yang (2009a) and Edwards et al. (2021), who used some identical conditions, followed by Zimmerman et al. (1993), who investigated correlated residuals, and other studies that are comparable in terms of explored coefficients.

The comparison is limited to findings for unidimensional conditions. Replications of previous findings and new insights are discerned. To avoid repetition, new insights were placed adjacent to the study they expanded upon the most.

**Listing 1.** Replicated and new insights from Study 1.

Green and Yang (2009a), Edwards et al. (2021)

*Replicated insights*

- α expectedly underestimates population reliability if τ-equivalence does not hold.
- α and ω performed similarly and that α even outperformed ω when population reliability,

sample size, and test length were low.

- $\omega$ slightly outperformed $\alpha$ in congeneric conditions.
- GLB mostly overestimates reliability.

*New insights*

- $\alpha$, followed by $\lambda_2$, underestimates reliability the most in nearly every congeneric condition.
- $\alpha$ outperforms some coefficients in terms of bias only if loadings are all relatively high (0.8) except for one item.
- $\omega_t$ was outperformed by $\lambda_4$ variants, $\mu_2$, KC, HB, $\rho$, and GLB variants in numerous congeneric conditions, especially with up to moderately high loadings or heterogeneous loadings with a range of low to high loadings.
- In congeneric conditions with highly heterogeneous loadings, coefficient performance does not converge with an increase in sample size and test length as in conditions in which $\tau$-equivalence holds or loadings are congeneric but all relatively high.
- If most loadings are homogeneous and high and sample size is large, coefficients perform quite similarly.

Zimmerman et al. (1993)

*Replicated insights*

- $\alpha$ expectedly underestimates population reliability if $\tau$-equivalence does not hold.
- $\alpha$ overestimates reliability when residuals are positively correlated.
- $\alpha$ can be imprecise with small sample sizes.

*New insights*

- $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, KC, HB, $\rho$, and GLB variants are similarly affected by correlated residuals as $\rho$ if all the residuals are mutually weakly positively correlated, regardless of the sample size.
- $\alpha$, $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, and $\omega_t$ tend to underestimate reliability even if there is residual correlation that is low and either positive or negative.
- $\omega_t$ and $\text{GLB}_A$ expectedly outperformed other coefficients, especially $\alpha$, if measurement model is among the ones used in Green and Yang (2009a) and all the residuals are weakly inter-correlated.
- KC, HB, and occasionally GLBA outperform other coefficients if some residual pairs are weakly correlated.
- When unidimensionality holds, factor loadings affect reliability coefficients the most, and higher-order interactions are trivial.

Osburn (2000)

*Replicated insights*

- $\lambda_{4(\text{max})}$, followed by GF, outperformed other coefficients, including $\alpha$ and $\lambda_2$, in unidimensional conditions.
- $\alpha$ underestimates reliability under severe violation of $\tau$-equivalence.
- Most coefficients, including $\lambda_2$, are mildly superior to $\alpha$ if the model is congeneric.

*New insights*

- $\lambda_{4(\text{max})}$ is not as superior to other coefficients.
- GF is among the least precise coefficients, but it can be useful with small sample sizes and relatively low loadings.
- Most coefficients are superior to $\alpha$ if the model is congeneric with up to moderate, or especially with highly heterogeneous loadings.

Thompson et al. (2010)

*Replicated insights*

- $\alpha$ performs similarly to other coefficients when $\tau$-equivalence holds.

- $\alpha$ was outperformed by other coefficients when $\tau$-equivalence was violated.
- $\alpha$ generally underestimates reliability and it can be severely biased with small sample sizes.

*New insights*

- $\lambda_{4(\max)}$ is not expectedly positively biased, even if residuals are weakly positively correlated.

Hunt and Bentler (2015)

*Replicated insights*

- $\lambda_{4(\max)}$ appears to be generally superior to $\lambda_{4(Q=0.05)}$ in terms of median bias but inferior in terms of precision unless loadings are up to moderate.
- $\text{GLB}_\text{A}$ can overestimate population reliability more than $\lambda_{4(\max)}$ and $\lambda_{4(Q=0.05)}$.
- $\lambda_{4(Q=0.05)}$ often outperforms $\alpha$.

*New insights*

- $\text{GLB}_\text{A}$ and especially $\text{GLB}_\text{M}$ are outperformed by $\omega_t$ more often than not since GLB variants tend to overestimate reliability.
- $\text{GLB}_\text{A}$ can outperform $\omega_t$ under various measurement models if the test is short and sample is large.
- $\text{GLB}_\text{A}$ does not always require large sample and approximately 200 can be sufficient, ANOVA showed loading magnitude is more relevant than sample size for $\text{GLB}_\text{A}$ bias.

Trizano-Hermosilla and Alvarado (2016)

*Replicated insights*

- $\alpha$ and $\omega$ converge in $\tau$-equivalent conditions.
- $\omega$ is superior to $\alpha$ in general, even with small sample sizes.
- GLB overestimates reliability under some conditions, especially if the sample is small.
- Coefficients are less biased in conditions with 12 items compared to conditions with 6 items.

*New insights*

- $\alpha$, $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, KC, HB, $\rho$, and GLB variants converge when $\tau$-equivalence holds and loadings are high, but if loadings are $\tau$-equivalent and equal to 0.5 or 0.2, coefficients differ and most of them tend to underestimate reliability.
- KC is a superior default choice to $\omega_t$.
- An increase in test length can result in improved precision with small sample sizes for all the coefficients.

Listing 1 shows that insights related to coefficients generally support the observation by Cho (2022) that no coefficient is appropriate for every condition. The fact that some insights are replicated further supports the conclusions from previous studies. However, the findings presented above have some limitations. More precisely, they do not apply to multidimensional structures, models with at least moderate residual correlation, ordinal scales, and data with extremely skewed items and outliers.

### 2.2. *Study 2*

*2.2.1. Conditions.* In Study 1, every condition represented unidimensional construct-relevant variance, while some conditions also included a certain degree of random multidimensionality in the form of correlated residuals. In the first scenario of Study 2, coefficient performance in conditions with fixed multidimensionality is examined more thoroughly. $\omega_h$ and $\omega_a$ are included since they converge to $\omega_t$ in Study 1 but differ from $\omega_t$ in multidimensional conditions. Conditions are therefore specified to cover correlated-factor models. Two-dimensional and three-dimensional structures are investigated since previous studies did not go further than two dimensions. Dimension correlation is specified to be 0.4, 0.55, and 0.7. Such levels were selected to represent conditions in which it is justifiable to estimate total reliability and cover moderate to strong correlation. Group factor loadings are specified as 0.2 and 0.5 to reflect low and medium loadings since higher loadings cannot be crossed with higher values of dimension correlation without obtaining a singular matrix. Sample size and test length levels are identical to those in Study 1 and crossed with every other level of remaining factors. The design is fully crossed and consists of 96 conditions.

In the second scenario, higher-order interactions of latent structure and some empirical factors are probed. Previous findings suggest that departures from non-normality and continuous scale result in reliability underestimation, but it was not investigated how these factors interact with the latent structure. Xiao and Hau (2023) crossed various non-normal distributions, factor loading levels, and sample sizes with both continuous and ordinal scales. However, these conditions were limited to six unidimensional items in which every item is identically distributed. Therefore, in the second scenario, sample size, test length, number of dimensions, factor loading, scale, and item distribution are varied simultaneously. Test length levels are taken over from the previous scenarios, but sample size levels are limited to 200 and 1000 since it is sufficient to explore interactions. As the number of dimensions and their correlation were examined more thoroughly in the first scenario, they are limited to one or two dimensions with correlation 0.55 selected as moderate correlation magnitude. Similarly, group factor loadings were limited to 0.3 and 0.6 to cover both relatively low loadings and moderate to high loadings. The scale was limited to continuous and ordinal five-point scales because the latter is commonly used in practice. Item distribution included levels in which every item is normally distributed, half of the items are normally distributed and half of the items are extremely asymmetrically distributed, and every item is extremely asymmetrically distributed. Extreme asymmetry was selected since it was shown that up to moderate asymmetry does not significantly affect reliability estimation, especially if the scale is continuous (Xiao & Hau, 2023). The design is fully crossed and consists of 96 conditions.

*2.2.2. Data generation.* Random normal variables with validity checks were generated like in Study 1. Function `sim.minor()` from `psych` package was used to obtain population inter-item correlation matrices for each number of dimensions, dimension correlation, and factor loading condition. Random variables were multiplied with the Cholesky decomposition of a particular population inter-item correlation matrix to obtain target latent structures. The data generation procedure was identical in every phase of both Study 2 scenarios, with the exception that distribution for target items was transformed into non-normal using Fleishman's (1978) power method and continuous scale was categorized into ordinal five-point scale in half conditions of the second scenario. Coefficients for the Fleishman polynomial are $a = -1$, $b = 0.45$, and $c = 0.12$. In the case of five-point ordinal scale and asymmetrical items, continuous scale was transformed into the target shape before categorization. Extreme asymmetry had skewness = 1.24 and kurtosis = 0.93 for continuous scale conditions, and skewness = 1.56 and kurtosis = 1.82 after discretization. Categorization was done using the equipercentile method. Benchmark was similarly defined as in Study 1 but is based on both general and group factor loadings and it is comparable to total reliability in Trizano-Hermosilla et al. (2021). Therefore, only the total reliability was specified as the benchmark value.

*2.2.3. Study 2 results.* Figure 6 illustrates coefficient performance pattern is generally similar over the conditions, while $\omega_h$ and $\omega_a$ differ compared to other coefficients. Precision improves as the $\lambda$ level, $r$, and $N$ increase. $\omega_h$ is expectedly the least biased if $r$ is up to 0.55

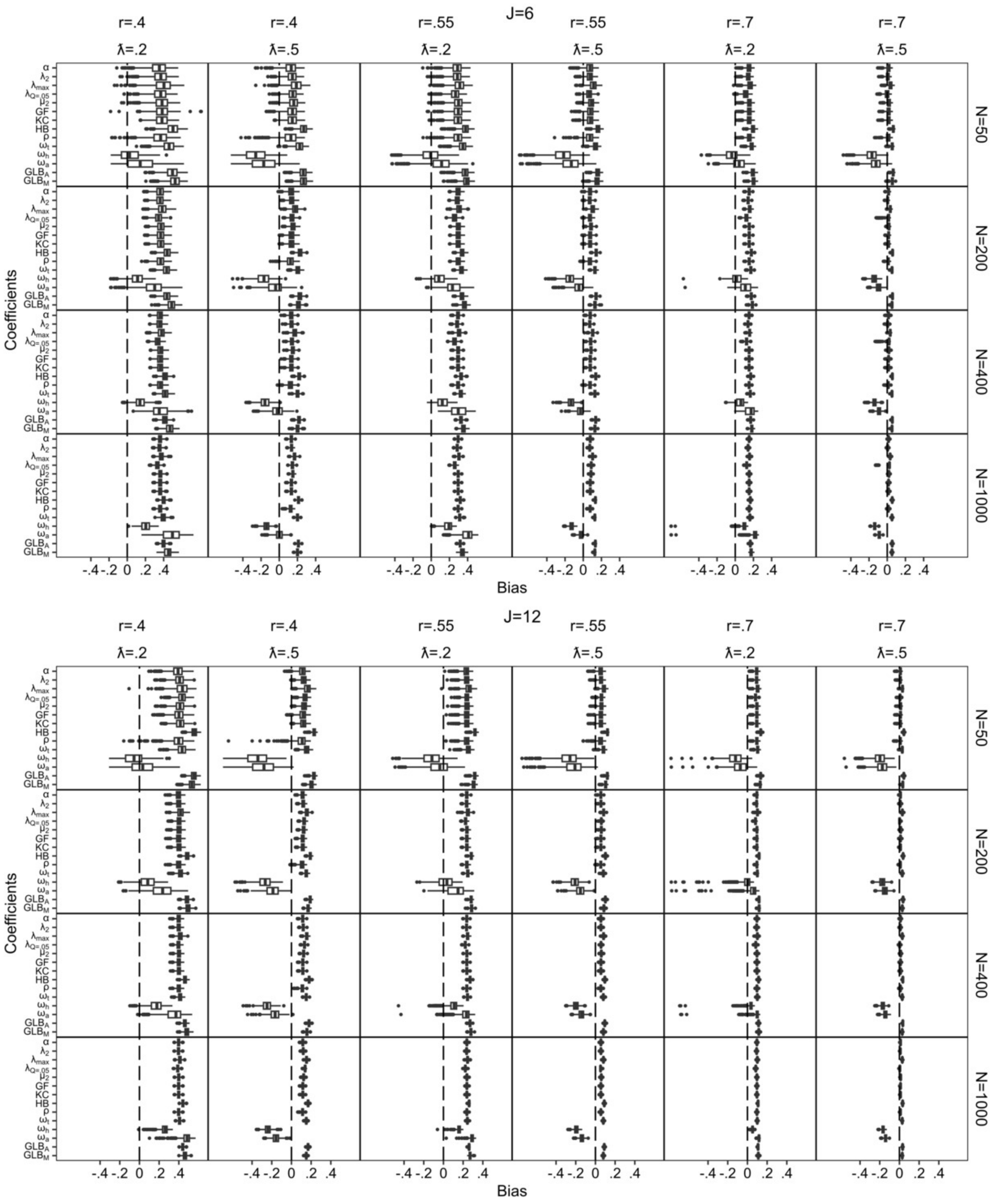

**Figure 6.** Two-dimensional conditions.

and loadings are low, but is the most imprecise in every condition. However, it outperformed $\omega_a$ in terms of bias. If $r$ is 0.7 and $\lambda$ is moderate, other coefficients outperform $\omega_h$ and $\omega_a$ and perform mutually similarly, while HB, GLB variants, and occasionally $\omega_t$ tend to slightly overestimate reliability. $\lambda_{4(Q=0.05)}$ is mostly trivially superior to other non-hierarchical coefficients, as well as GF if $J$ is 12. The coefficient performance pattern is quite similar in three-dimensional conditions and the results are displayed in Figure A3. The main difference is that, compared to two-dimensional conditions, $\omega_a$ outperforms $\omega_h$ if loadings are moderate.

In both two-dimensional and three-dimensional conditions, an increase in $r$ and loadings results in improved precision, especially if combined with large $N$, regardless of dimensionality. It also appears an increase in test length reduces precision when unidimensionality does not hold, regardless of $r$ and the number of dimensions, except if loadings are moderate and $r$ is 0.7.

Potential interactions are probed using ANOVAs with coefficient bias as a dependent variable using partial $\eta^2$ as the effect size estimate. Every level of each factor was included. Effect sizes were treated like in the previous ANOVA. The results are displayed in Figure A4.

There are a few practically relevant insights from this ANOVA. $\lambda$ is the most influential, followed by $r$. The interaction $\lambda \times r$ is non-trivial, while higher-order interactions have limited relevance in this combination of factors. $J$ and $N$ have mostly trivial effects, but they are relevant in specific interactions. Overall, $\lambda$ and $r$ are of primary importance for coefficient selection. Regarding specific coefficients, $\lambda$ has a significant impact on all coefficients, affecting KC the most and $\lambda_{4(\max)}$ the least. The effect of $r$ is non-trivial for all coefficients except $\omega_h$. Among the interactions, $\lambda \times r$ influences GF, $\omega_h$, and $\omega_a$ the least, while the other coefficients are affected similarly. Other interaction effects have limited practical utility, with only a subset of coefficients showing non-trivial effects.

Overall, the influence of particular factors on the bias of reliability coefficients appears to be more intricate if unidimensionality does not hold. However, these findings still assume normal distribution and continuous scale in every condition, which is not the case in scenario 2. Due to expected differences in coefficient performance resulting from the influence of dimensionality, results are displayed separately for unidimensional and two-dimensional conditions. The results of scenario 2 for unidimensional conditions are displayed in Figure 7.

Figure 7 shows coefficients perform similarly if all the items are normally distributed, even if the scale is ordinal with underlying normal distribution (D), except for $\text{GLB}_\text{M}$, which tends to overestimate reliability. Coefficients are generally affected by asymmetrical D. If the scale is ordinal and the underlying D is asymmetrical, coefficients additionally underestimate reliability compared to conditions with continuous scale. However, like in Study 1, in which every condition assumes continuous D, estimates converge with an increase in $N$ and factor loadings with ordinal scales as well, if the underlying D is normal. If all the items are asymmetrical, coefficients expectedly systematically underestimate reliability.

$\text{GLB}_\text{M}$ seems to be the least biased if all the items are asymmetrical and if also $N$ is large ($N = 1000$), while $\text{GLB}_\text{A}$ is the least biased if also $N$ is not large ($N = 200$). However, HB performs equally well as GLB variants if all items are asymmetrical, loadings are low, and $N$ is not large. However, if half the items are asymmetrical and half are normally distributed, there are more notable differences in coefficient performance. In such cases, for six-item conditions, KC is the least biased overall and equalizes with HB if $\lambda$ is not low, while $\text{GLB}_\text{M}$ performs nearly like KC if $\lambda$ is low and the scale is continuous. Conversely, $\alpha$ appears to be affected by result asymmetry the most overall.

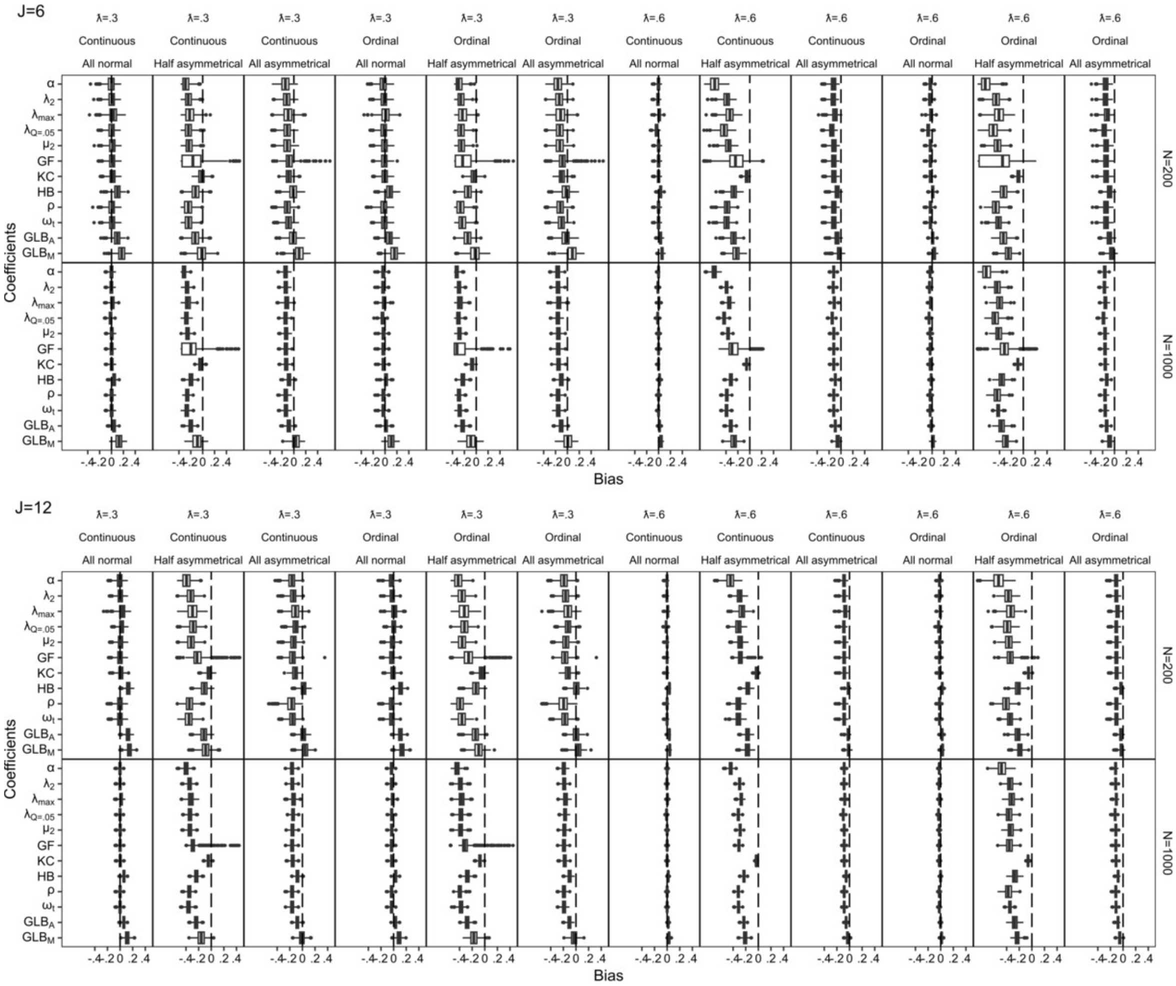

**Figure 7.** Unidimensional conditions with distribution asymmetry and ordinal scale.

Finally, results from Figure 8 with continuous scale and normal distribution are comparable to conditions in Figure 1 because τ-equivalence holds in both. Their comparison reveals that coefficients begin to perform similarly if loadings are at least 0.6. However, this is limited to unidimensional τ-equivalent conditions with large samples and normally distributed results. The results of scenario 2 for two-dimensional conditions are displayed in Figure 8.

Figure 8 highlights the distinctions between conditions where half of the items are asymmetrical and conditions where all the items are asymmetrical. These distinctions are more pronounced in two-dimensional conditions than in the unidimensional conditions depicted in Figure 7. Moreover, coefficients perform relatively similarly if all the items are symmetrical and if all the items are asymmetrical. When all the items in the dataset follow a symmetrical distribution and have low loadings, coefficients tend to overestimate reliability. Conversely, when loadings are not low, coefficients predominantly tend to underestimate reliability.

If loadings are low and items are all symmetrical or all asymmetrical, coefficients perform similarly, even more similarly in 12-item conditions. In such conditions with all symmetrical items, $\lambda_{4(Q=0.05)}$ is the least biased, and with all asymmetrical items, $\lambda_{4(Q=0.05)}$ is again the least biased if the scale is continuous, while most other coefficients are relatively unbiased if the scale is ordinal. If loadings are low and half the items are asymmetrical, KC is the least biased, closely followed by HB and $\text{GLB}_A$ in 12-item conditions. If loadings are not

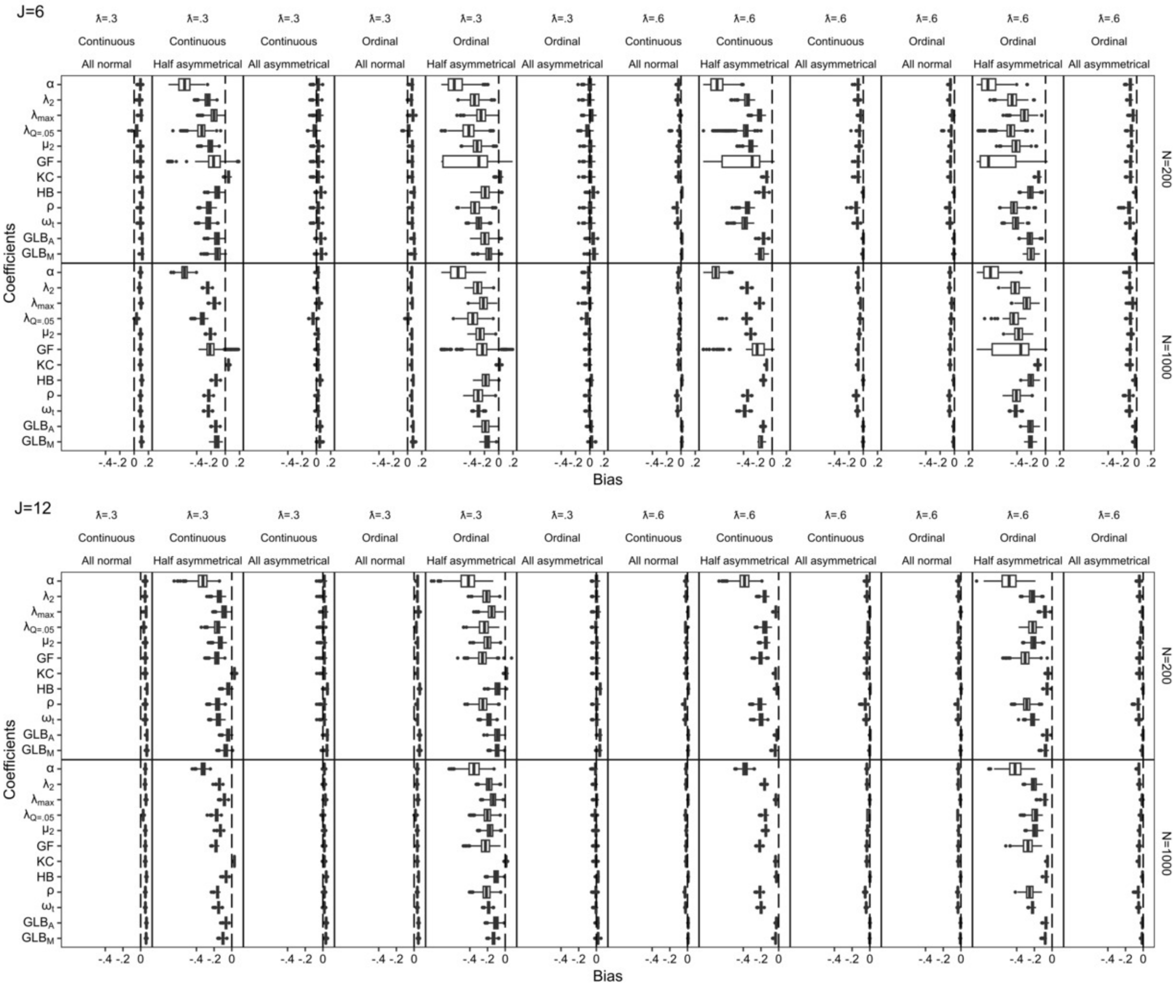

**Figure 8.** Two-dimensional conditions with distribution asymmetry and ordinal scale.

low and items are all symmetrical or all asymmetrical, HB and GLB variants are the least biased. If loadings are not low and half the items are asymmetrical, KC is the least biased in six-item conditions, and $\lambda_{4(\max)}$, KC, HB, and GLB variants are the least biased in 12-item conditions.

Based on ANOVA from the previous scenario, it was demonstrated interactions are intricate when unidimensionality does not hold (see Figure A4). ANOVA in this scenario suggests interactions are even more intricate if distribution and scale are also considered. In Figure A5, interactions among $N$, $J$, result D, scale (S), $\lambda$, and number of dimensions/factors $f$ are displayed. Every level of each factor was retained and it should be noted that D has three levels, as displayed in Figure 6 and Figure A3. Effect sizes were treated like in previous ANOVAs.

ANOVA in this scenario revealed that D affects bias of reliability coefficients the most among the main effects, followed by $f$, $\lambda$, S, and $J$. This effect is limited to extreme asymmetry. In terms of interactions, $f \times \mathrm{D}$, $\lambda \times f$, $J \times f \times \mathrm{D}$, and $J \times \mathrm{D}$ affect bias more than S and $J$, while the former is the most relevant interaction effect. The main effect of D is generally large, and affects $\alpha$ the most, like in unidimensional conditions. The main effect of $f$ is mostly large but trivial for GF and $\mathrm{GLB}_\mathrm{A}$. The main effect of $\lambda$ is large for nearly every coefficient and medium for $\lambda_{4(\max)}$ and GF, but generally smaller than the main effect of D. The main effect of S is trivial for $\alpha$, GF, and $\mathrm{GLB}_\mathrm{A}$. There are no non-trivial interactions involving S. However, there are several interactions involving D, $\lambda$, $f$, and $J$, especially the former. Higher-order interactions are trivial in the context of latent structure and some empirical factors that influence bias of reliability coefficients.

*2.2.4. Study 2 findings in the context of previous research.* The findings of Study 2 are contextualized in Listing 2, which summarizes insights provided in Figures 8 and A1–A4. The insights are compared to findings from previous studies.

Unlike Listing 1, these insights focus on findings in conditions with multidimensional structures, extreme skewness, and ordinal scales. Therefore, some studies from Listing 1 appear in Listing 2 as well. The process of comparing insights begins with studies that investigated coefficients in multidimensional structures and then proceeds to studies that addressed issues related to skewness and ordinal scales.

**Listing 2.** Replicated and new insights from Study 2.

Trizano-Hermosilla et al. (2021)

*Replicated insights*

- $\omega_t$, $\mathrm{GLB}_\mathrm{A}$, and $\mathrm{GLB}_\mathrm{M}$ are less biased reliability estimates in unidimensional structures compared to $\omega_h$ and $\omega_a$.

*New insights*

- $\omega_h$ is expectedly the least biased in two-dimensional conditions if factor correlation is up to moderate and loadings are low.
- $\omega_a$ outperforms $\omega_h$ in three-dimensional conditions with moderate loadings.
- $\alpha$, $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, and KC are occasionally more useful than $\omega_h$ and $\omega_a$ if the factor correlation is high and loadings are moderate.
- Precision of all the coefficients improves with an increase in loadings, factor correlation, and sample size, but decreases with an increase in test length.
- ANOVA confirmed $\omega_h$ and $\omega_a$ are less affected by the factor correlation than other coefficients, but they are similarly affected by loadings.
- ANOVA suggests loadings and factor correlation are the most relevant in selecting the

| |
|---|
| appropriate coefficient for particular conditions. |
| Osburn (2000) |
| *Replicated insights*<br>• $\lambda_{4(\max)}$ is the least biased in two- and three-dimensional conditions with parallel, $\tau$-equivalent, and congeneric model, and low to high correlation between factors.<br>*New insights*<br>• Among the non-hierarchical coefficients, $\lambda_{4(Q=0.05)}$ is mostly trivially superior to $\lambda_{4(\max)}$, and GF is trivially superior to $\lambda_{4(\max)}$ if test length is 12.<br>• $\omega_h$ and $\omega_a$ are superior to non-hierarchical coefficients when factors are highly correlated. |
| Hunt and Bentler (2015) |
| *Replicated insights*<br>• $\alpha$ substantially underestimates reliability.<br>• $\lambda_{4(Q=0.05)}$ is useful for two-dimensional conditions when $\tau$-equivalence both does or does not hold and for all sample sizes.<br>• $\lambda_{4(Q=0.05)}$ is superior to $\alpha$ in two-dimensional conditions.<br>• $\lambda_{4(\max)}$ and $\text{GLB}_\text{A}$ overestimate reliability.<br>*New insights*<br>• $\lambda_{4(\max)}$ and $\text{GLB}_\text{A}$ overestimate reliability, unless factor correlation is high and loadings are moderate, regardless of sample size and test length.<br>• $\lambda_{4(Q=0.05)}$ is the least biased in two-dimensional conditions with moderately correlated factors if all loadings are low, regardless of the scale and sample size.<br>• $\lambda_{4(Q=0.05)}$ is the least biased if all the items are asymmetrical and sample size is 200 and outperformed $\lambda_{4(\max)}$ in two-dimensional conditions with moderately correlated factors, regardless of the scale and sample size. |
| Thompson et al. (2010) |
| *Replicated insights*<br>• $\lambda_{4(\max)}$ outperforms $\alpha$ in two-dimensional conditions.<br>*New insights*<br>• $\lambda_{4(\max)}$ is not the least biased and the most precise in two-dimensional conditions, but similarly biased and precise as other non-hierarchical coefficients.<br>• KC is the most precise coefficient generally. |
| Trizano-Hermosilla and Alvarado (2016) |
| *Replicated insights*<br>• $\omega_t$ slightly outperforms $\alpha$ in terms of bias and GLB variants are positively biased if items are normally distributed regardless of measurement model and sample size.<br>• $\alpha$ and $\omega_t$ underestimate reliability if items are asymmetrical regardless of measurement model and sample size.<br>• GLB variants are relatively robust against extreme asymmetry if all items are asymmetrical and outperform $\alpha$ and $\omega_t$ in such conditions.<br>*New insights*<br>• $\alpha$ is the least robust against extreme asymmetry.<br>• HB is the most useful and outperforms GLB variants if all the items are asymmetrical and loadings are not high.<br>• KC, and occasionally HB, outperform GLB variants if half the items are extremely asymmetrical.<br>• $\alpha$, $\lambda_2$, $\lambda_4$ variants, $\mu_2$, GF, and $\rho$ underestimate reliability less if all the items are extremely asymmetrical than if half the items are extremely asymmetrical. |

- Performance pattern of all the coefficients is more similar in conditions in which all the items are normally distributed to conditions in which all the items are asymmetrical than to conditions in which half the items are asymmetrical.

Xiao and Hau (2023)

*Replicated insights*

- Coefficients are biased in unidimensional conditions if factor loadings are moderate.
- High loadings and more scale points generally mitigate the biasing effect of extremely asymmetrical items.
- Coefficients are differently affected by factors that influence reliability estimation.
- GLB is robust against asymmetry, but α and ω-family coefficients are not.

*New insights*

- If item asymmetry is extreme, distribution influences bias of reliability coefficients more compared to number of dimensions and factor loadings, number of dimensions is generally more influential compared to factor loadings, while the main effects of sample size, test length, and scale are in such cases trivial for all the coefficients.
- The interaction of distribution with number of dimensions appears to be more relevant for reliability estimation bias than the interaction of distribution with scale if unidimensionality does not hold.

Listing 2 shows that, similarly to Listing 1, some insights are replicated, and that further supports the conclusions from previous studies, while some insights are new and potentially useful for practical reliability estimation. The shared limitation with Study 1 is that these insights do not apply to data with outliers.

## 3. General discussion

Some joint insights from Study 1 and Study 2 expand the existing analytical theory. For instance, Raykov (1997) showed analytically that α is negligibly biased beyond test length four or six if loadings are at least 0.6. As indicated by Figures 1 and A3, if unidimensionality and τ-equivalence hold, loadings are at least 0.6, the scale is either continuous or five-point ordinal, sample is large, and the results are normally distributed, α performs similarly to other coefficients. In such cases, it makes little difference which coefficient is used. Additional support for the existing analytical theory is that not only α, but most other coefficients are relatively unbiased and precise in such conditions. Differences in coefficient performance are mostly observed if the model is τ-equivalent with lower loadings or the model is congeneric, as shown in Study 1. Moreover, in Study 1, it was demonstrated that even when all loadings are high ($\lambda = 0.8$) and one item violates τ-equivalence, α does not exhibit the highest level of bias among the coefficients. These findings expand existing analytical theory (Raykov, 1997, 1998), which suggests that the bias of α can be substantial even if one item violates τ-equivalence.

Moreover, Study 1 and Study 2 generally support the criticism of the indiscriminate use of α (e.g., Cho, 2021b; Green & Yang, 2009a; Sijtsma, 2009; Sijtsma & Pfadt, 2021). These studies, however, also highlight that α can be useful in specific conditions (Raykov & Marcoulides, 2019). McNeish (2018) concluded α is outperformed by every other coefficient. Study 1 and Study 2 demonstrated it is mostly so, but not universally, as observed by Cho (2022). In limited conditions in which α is not outperformed by every other coefficient, it is never the least biased. As Bentler (2021) stated, α is simply a conservative lower bound to reliability in some cases. α is theoretically relevant, but there are more useful alternatives for practical reliability estimation.

Previously, it was suggested $\alpha$ can be supplanted with $\omega_t$ (e.g., Revelle & Zinbarg, 2009), which would result in at least a slight improvement in most situations according to Study 1, Study 2, and some previous research. Moreover, findings in Study 1 and Study 2 supported the suggestion that $\lambda_2$, $\mu_2$, and GF are generally more appropriate than $\alpha$ (Cho, 2021b) in terms of bias. However, these coefficients are also occasionally outperformed by $\lambda_4$ variants, KC, HB, $\rho$, $\omega_t$, GLB variants, or even $\omega_h$ and $\omega_a$ in some congeneric and multidimensional conditions. HB, KC, and $\text{GLB}_\text{A}$ were found to be more useful than previously considered.

However, it should be noted that the appropriateness of these coefficients was evaluated primarily in terms of bias. There are conditions in which coefficient precision can be generally questionable, such as those with low loadings and small sample sizes. This issue is more prominent with some coefficients, such as GLB variants, especially $\text{GLB}_\text{M}$, as well as GF. Thus, while the findings of this study can inform coefficient selection, it is advisable to consider using multiple coefficients for reliability estimation due to their imprecision. Moreover, the findings are applicable for test length up to 12 items and general tendency toward reliability overestimation by GLB variants (e.g., Revelle & Zinbarg, 2009) should be taken into account. Key takeaways for practical reliability estimation are provided in Table 1.

These findings also revealed gaps for future research. There are no coefficients that are precise in conditions with small sample sizes and up to moderate loadings, which was evaluated using bias distribution. It might be useful to use additional performance measures in future studies, such as (root) mean square error, to get additional insights about coefficient accuracy. Moreover, while this study has compared this set of coefficients in numerous conditions, further extensions are still possible. For instance, the effect of outliers and missing data were not included in the designs. Also, while the test lengths used in the designs are reasonable for unidimensional constructs, multidimensional conditions can be investigated using a larger number of items. Regarding the latent structure, future research might focus more on congeneric models where loadings are not high, as well as multidimensional structures and structures with highly correlated residual pairs.

**Table 1.** Recommended coefficients for specific conditions.

| Condition | Coefficient |
|---|---|
| Conditions in which it makes little difference which coefficient is used | |
| Unidimensional, $\tau$-equivalent loadings that are at least 0.6, uncorrelated residuals, large sample, all items normally distributed | / |
| Conditions in which coefficients can differ in performance non-trivially | |
| *Unidimensional, $\tau$-equivalent loadings, uncorrelated residuals, and normal item distribution* | |
| Low loadings, small sample, short test | $\rho$ |
| Low loadings, small sample, test not short | $\mu_2$ |
| Moderate loadings, small sample, short test | $\mu_2$ |
| Moderate loadings, small sample, test not short | $\lambda_{4(\text{max})}$ |
| Moderate loadings, moderate sample size | HB |
| Moderate loadings, large sample | $\text{GLB}_\text{M}$ |

*Continued on next page*

**Table 1.** Recommended coefficients for specific conditions. (Continued)

| Condition | Coefficient |
|---|---|
| *Unidimensional, congeneric loadings, uncorrelated residuals, and normal item distribution* | |
| Mostly low loadings, small sample size | $\mu_2$, $\lambda_{4(Q=0.05)}$ |
| Mostly low loadings, at least moderate sample size | $\rho$ |
| Moderate loadings, at least moderate sample size | HB, $\text{GLB}_\text{A}$ |
| Heterogeneous loadings of any magnitude, small sample size | $\lambda_{4(\text{max})}$, $\lambda_{4(Q=0.05)}$ |
| Approximately half low-half moderate loadings, small sample size | $\lambda_{4(\text{max})}$, $\lambda_{4(Q=0.05)}$ |
| Heterogeneous loadings of any magnitude, at least moderate sample size | KC |
| Approximately half low-half moderate loadings, at least moderate sample size | KC |
| *Unidimensional, weakly correlated residuals, and normal item distribution* | |
| Every residual pair correlated | KC, HB, $\text{GLB}_\text{A}$ |
| Proportions of correlated residuals | $\text{GLB}_\text{M}$, $\text{GLB}_\text{A}$, $\lambda_{4(Q=0.05)}$ |
| *Unidimensional, positively correlated residuals, and extremely asymmetrical item distribution* | |
| Low loadings, all the items are extremely asymmetrical | KC, HB |
| High loadings, all the items are extremely asymmetrical | $\text{GLB}_\text{M}$, $\text{GLB}_\text{A}$ |
| Half the items are extremely asymmetrical | KC |
| *Two-dimensional, extremely asymmetrical items* | |
| Half the items are extremely asymmetrical | KC |
| High loadings, all the items extremely asymmetrical | $\text{GLB}_\text{M}$, $\text{GLB}_\text{A}$, HB |
| Up to moderate loadings, all the items are extremely asymmetrical | $\lambda_{4(Q=0.05)}$ |
| *Multidimensional conditions* | |
| Two dimensions, low loadings | $\omega_h$, $\omega_a$ |
| Three dimensions, low loadings | $\omega_a$ |
| At least moderate loadings | Any non-hierarchical coefficient, $\lambda_{4(Q=0.05)}$ slightly preferable over others |

## References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*(2), 155–173. https://doi.org/10.1007/bf02294170

Bentler, P. M. (2021). Alpha, FACTT, and beyond. *Psychometrika*, *86*(4), 861–868. https://doi.org/10.1007/s11336-021-09797-8

Caeiro, F., & Mateus, A. (2022). *randtests: Testing randomness in R* (Version 1.0.1) [Computer software]. The Comprehensive R Archive Network. https://cran.r-project.org/package=randtests

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

Cho, E. (2021a). *eunscho/unirel: Compute and compare unidimensional reliability coefficients* (Version 1.2.0) [Computer software]. R Package Documentation. https://rdrr.io/github/eunscho/unirel

Cho, E. (2021b). Neither Cronbach's alpha nor McDonald's omega: A commentary on Sijtsma and Pfadt. *Psychometrika*, *86*(4), 877–886. https://doi.org/10.1007/s11336-021-09801-1

Cho, E. (2022). The accuracy of reliability coefficients: A reanalysis of existing simulations. *Psychological Methods*. https://doi.org/10.1037/met0000475

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

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement*, *81*(6), 1089–1117. https://doi.org/10.1177/0013164421994184

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521–532. https://doi.org/10.1007/bf02293811

Gessaroli, M., & Champlain, A. F. (2005). Test dimensionality: Assessment of. In *Everitt, B. S. & Howell, D. (Eds.). Encyclopedia of Statistics in Behavioral Science* (pp. 2014–2021). John Wiley & Sons.

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika*, *48*(1), 99–111. https://doi.org/10.1007/bf02314679

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

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

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, *2*, 104–129. https://doi.org/10.2307/270785

Hunt, T. D. (2019). *Lambda4: Collection of internal consistency reliability coefficients* (Version 3.0) [Computer software]. R Package Documentation. https://rdrr.io/cran/Lambda4

Hunt, T. D., & Bentler, P. M. (2015). Quantile lower bounds to reliability based on locally optimal splits. *Psychometrika*, *80*(1), 182–195. https://doi.org/10.1007/s11336-013-9393-6

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*(4), 567–578. https://doi.org/10.1007/bf02295979

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. https://doi.org/10.1007/bf02291393

Kaiser, H. F., & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika*, *30*(1), 1–14. https://doi.org/10.1007/bf02289743

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

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989x.4.1.84

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100–117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Taylor & Francis. https://doi.org/10.4324/9781410601087

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Moltner, A., & Revelle, W. (2020). R: Find the greatest lower bound to reliability. http://personality-project.org/r/psych/help/glb.algebraic.html

Novak, J., & Rebernjak, B. (2023). There are many greater lower bounds than cronbach's α: A monte carlo simulation study. *Measurement: Interdisciplinary Research and Perspectives*, *21*(1), 1–28. https://doi.org/10.1080/15366367.2022.2031484

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*(1), 1–13. https://doi.org/10.1007/bf02289400

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*(3), 343–355. https://doi.org/10.1037/1082-989x.5.3.343

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Raykov, T. (1997). Scale reliability, cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*(4), 329–353. https://doi.org/10.1207/s15327906mbr3204_2

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*(4), 375–385. https://doi.org/10.1177/014662169802200407

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, *79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Revelle, W. (2022). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.2.5) [Computer software]. The Comprehensive R Archive Network. https://cran.r-project.org/package=psych

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

Santos Fernandez, E. (2014). *Johnson: Johnson transformation* (Version 1.4) [Computer software]. R Package Documentation. https://rdrr.io/cran/Johnson

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

Shapiro, A., & ten Berge, J. M. F. (2002). Statistical inference of minimum rank factor analysis. *Psychometrika*, *67*(1), 79–94. https://doi.org/10.1007/bf02294710

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, *53*(5), 676–694. https://doi.org/10.1080/00273171.2018.1476221

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

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Strimmer, K., Strimmer, T., Jendoubi, Kessy, A., & Lewin, A. (2022). *whitening: Whitening and high-dimensional canonical correlation analysis* (Version 1.4.0) [Computer software]. The Comprehensive R Archive Network. https://cran.r-project.org/package=whitening

ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*(4), 575–579. https://doi.org/10.1007/bf02293815

Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the maximal split-half coefficient to estimate reliability. *Educational and Psychological Measurement*, *70*(2), 232–251. https://doi.org/10.1177/0013164409355688

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology*, *7*, 769. https://doi.org/10.3389/fpsyg.2016.00769

Trizano-Hermosilla, I., Gálvez-Nieto, J. L., Alvarado, J. M., Saiz, J. L., & Salvo-Garrido, S. (2021). Reliability estimation in multidimensional scales: Comparing the bias of six estimators in measures with a bifactor structure. *Frontiers in Psychology*, *12*, 508287. https://doi.org/10.3389/fpsyg.2021.508287

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22–29. https://doi.org/10.1111/j.1745-3992.1996.tb00803.x

Wickham, H. (2016). Data analysis. In *Ggplot2* (pp. 189–201). Springer. https://doi.org/10.1007/978-3-319-24277-4_9

Wilke, C. O. (2022). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'* (Version 1.1.0) [Computer software]. The Comprehensive R Archive Network. https://cran.r-project.org/package=cowplot

Xiao, L., & Hau, K.-T. (2023). Performance of coefficient alpha and its alternatives: Effects of different types of non-normality. *Educational and Psychological Measurement*, *83*(1), 5–27. https://doi.org/10.1177/00131644221088240

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, *53*(1), 33–49. https://doi.org/10.1177/0013164493053001003

## Appendix A Supplementary material

**Listing A1.** Design for a specific scenario.

Study 1, Scenario 1

*Factors and their levels*

- Population loadings for 6-item and 12-item test (Green & Yang, 2009b)
- Sample size (50, 200, 400, 1000)
- Residual correlation (0, 0.05, 0.10).

*Design type*

- Partially crossed
- $14 \times 4 \times 3 + 14 \times 4 \times 3 = 336$

Study 1, Scenario 2

*Factors and their levels*

- Population loadings ($\lambda = 0.2$, $\lambda = 0.5$, $\lambda = 0.8$)
- Test length (6, 12)
- Sample size (50, 200, 400, 1000)
- Correlated residual pairs (one pair, two pairs)
- Residual correlation (0.05, 0.15)
- Residual correlation direction (positive, negative)

*Design type*

- Fully crossed
- $3 \times 2 \times 4 \times 2 \times 2 \times 2 = 192$

Study 2, Scenario 1

*Factors and their levels*

- Population group factor loadings ($\lambda = 0.2$, $\lambda = 0.5$)
- Number of factors (2, 3)
- Factor correlation (0.40, 0.55, 0.70)
- Sample size (50, 200, 400, 1000)
- Test length (6, 12)

*Design type*

- Fully crossed
- $2 \times 2 \times 3 \times 4 \times 2 = 96$

Study 2, Scenario 2

*Factors and their levels*

- Population group factor loadings ($\lambda = 0.3$, $\lambda = 0.6$)
- Number of factors (1, 2)
- Sample size (200, 1000)
- Test length (6, 12)
- Scale (continuous, five-point ordinal)
- Item distribution (symmetrical, extremely asymmetrical)
- Proportion of extremely asymmetrical items (none, half, all)

*Design type*

- Fully crossed
- $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 3 = 192$

**Table A1.** Missing values for particular coefficients across the scenarios.[5]

| | Study 1 | | Study 2 | |
|---|---|---|---|---|
| Coefficient | Scenario 1 | Scenario 2 | Scenario 1 | Scenario 2 |
| $\alpha$ | 1621 (0.0048) | 3219 (0.0168) | 2849 (0.0297) | 3244 (0.0169) |
| $\lambda_2$ | 440 (0.0013) | 1017 (0.0053) | 940 (0.0010) | 297 (0.0015) |
| $\lambda_{4(\max)}$ | 1042 (0.0031) | 1833 (0.0095) | 1757 (0.0183) | 528 (0.0028) |
| $\lambda_{4(Q=0.05)}$ | 313 (0.0009) | 666 (0.0034) | 616 (0.0064) | 282 (0.0015) |
| $\mu_2$ | 363 (0.0011) | 848 (0.0044) | 794 (0.0083) | 234 (0.0012) |
| GF | 8748 (0.0260) | 11 442 (0.0596) | 8071 (0.0841) | 9271 (0.0483) |
| KC | 0 | 0 | 0 | 0 |
| HB | 0 | 0 | 0 | 0 |
| $\rho$ | 0 | 0 | 0 | 0 |
| $\omega_t$ | 286 (0.0009) | 768 (0.0040) | 63 (0.0007) | 248 (0.0013) |
| $\text{GLB}_\text{A}$ | 141 (0.0004) | 95 (0.0005) | 0 | 64 (0.0003) |
| $\text{GLB}_\text{M}$ | 0 | 0 | 286 (0.0030) | 0 |
| $\omega_a$ | — | — | 1 (0.0000) | — |
| $\omega_h$ | — | — | 55 (0.0006) | — |

**Table A2.** Population models from Green and Yang (2009a).

| 6-item conditions | | | | | | 12-item conditions | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 |
| 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| 0.8 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.8 | 0.8 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.8 | 0.8 | 0.2 | 0.2 | 0.2 | 0.2 | 0.8 | 0.8 | 0.8 | 0.8 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.8 | 0.8 | 0.8 | 0.2 | 0.2 | 0.2 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.8 | 0.8 | 0.8 | 0.8 | 0.2 | 0.2 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.2 | 0.2 | 0.2 | 0.2 |
| 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.2 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.2 | 0.2 |
| 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| 0.8 | 0.8 | 0.5 | 0.5 | 0.2 | 0.2 | 0.8 | 0.8 | 0.8 | 0.8 | 0.5 | 0.5 | 0.5 | 0.5 | 0.2 | 0.2 | 0.2 | 0.2 |

[5] The table provides a comprehensive overview of coefficients and their associated proportions of missing values across the scenarios. It is evident that the proportion of missing values is consistently low for all coefficients, whereas KC, HB, and $\rho$ have no missing values.

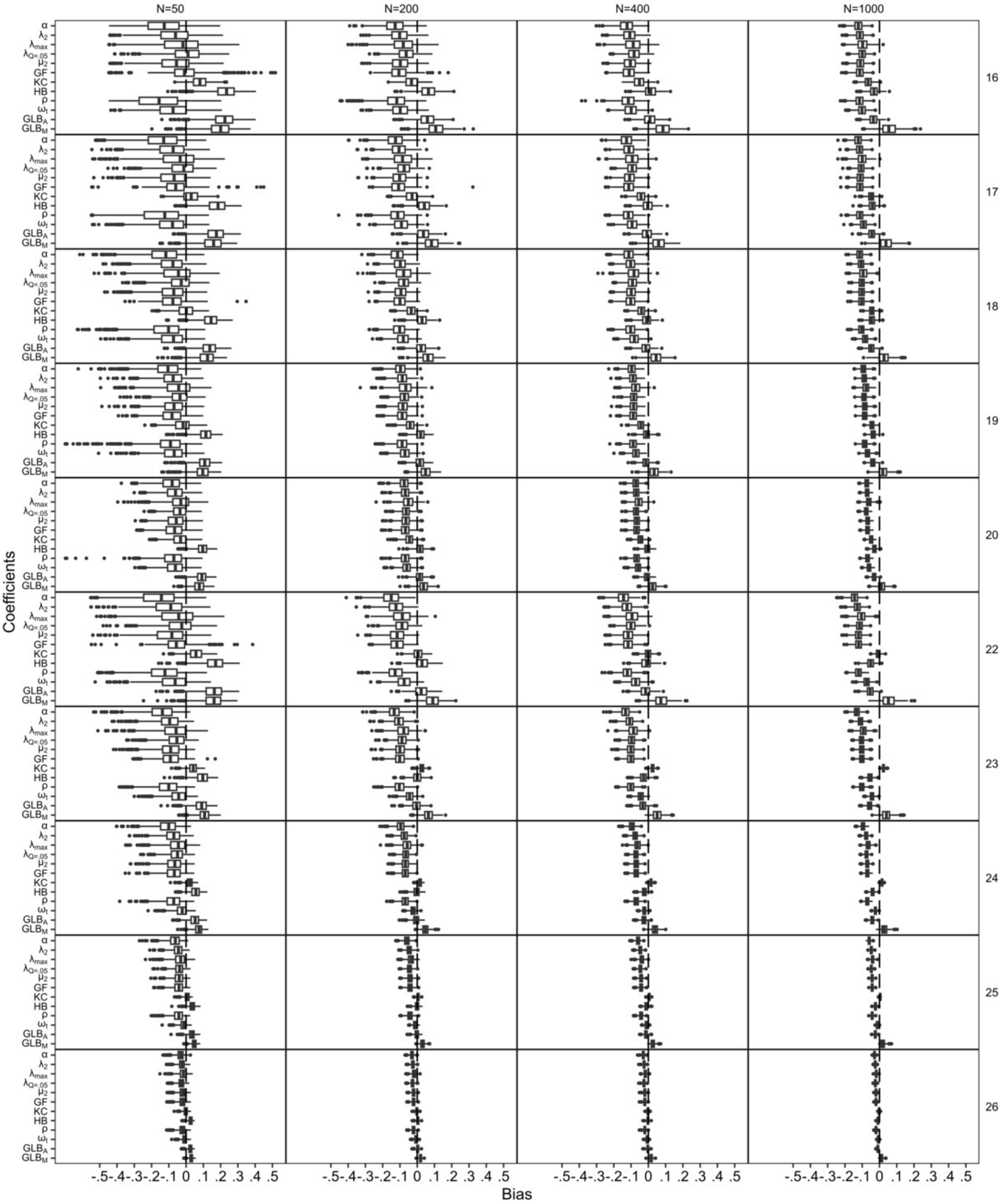

**Figure A1.** 12-item conditions with an increasing violation of τ-equivalence.

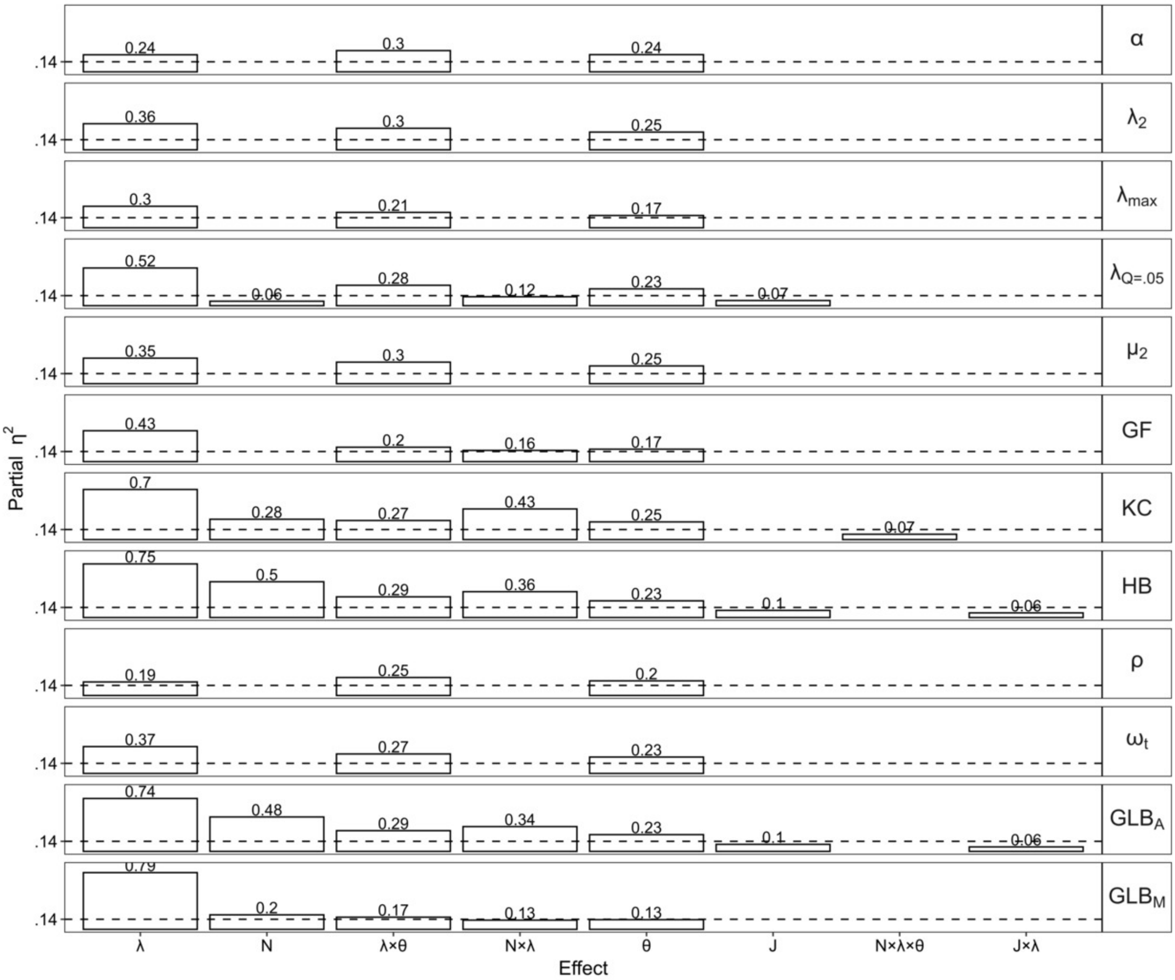

**Figure A2.** Effect sizes for nontrivial main and interaction effects.

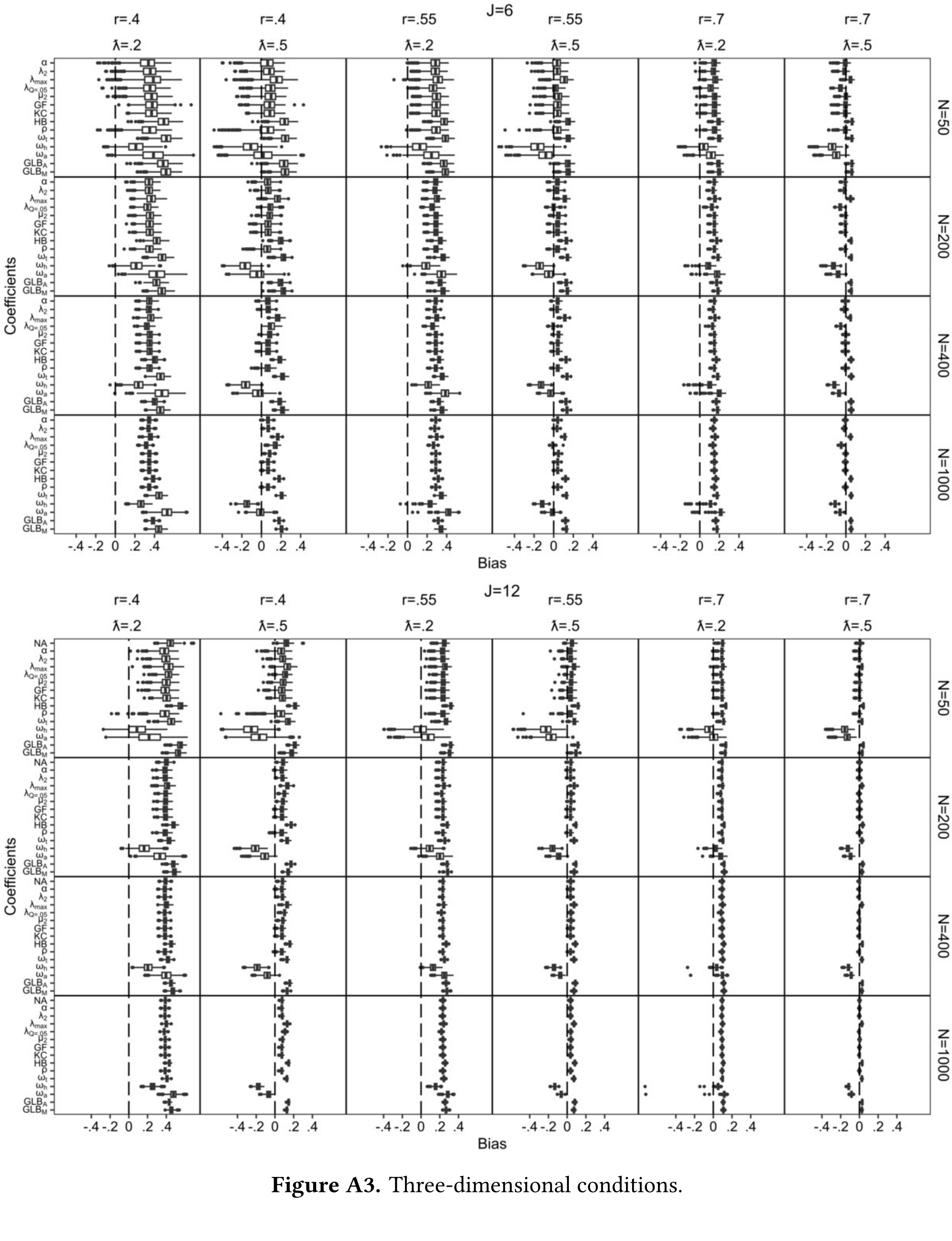

**Figure A3.** Three-dimensional conditions.

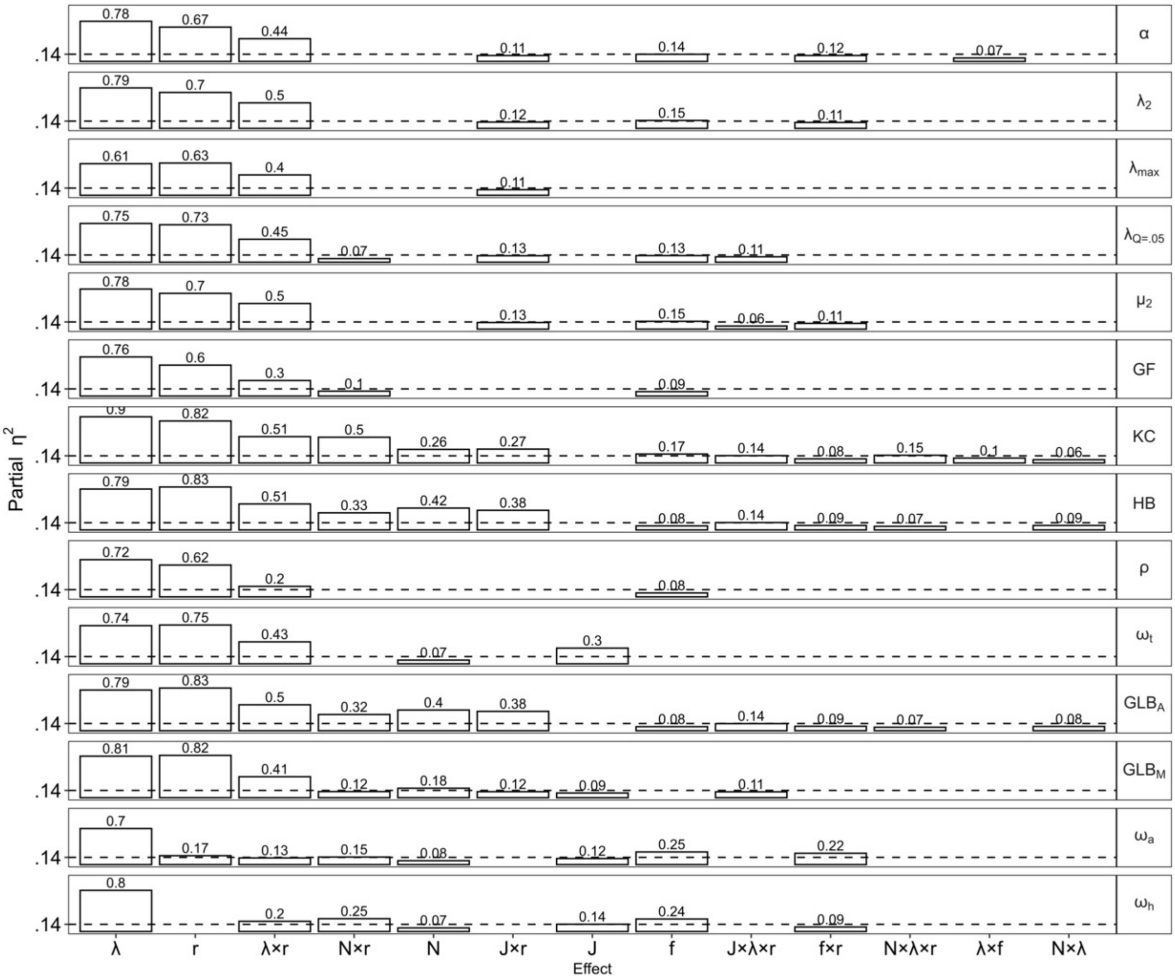

**Figure A4.** Effect sizes of particular factors and nontrivial interactions per coefficient for multidimensional conditions with normal distribution and continuous scale.

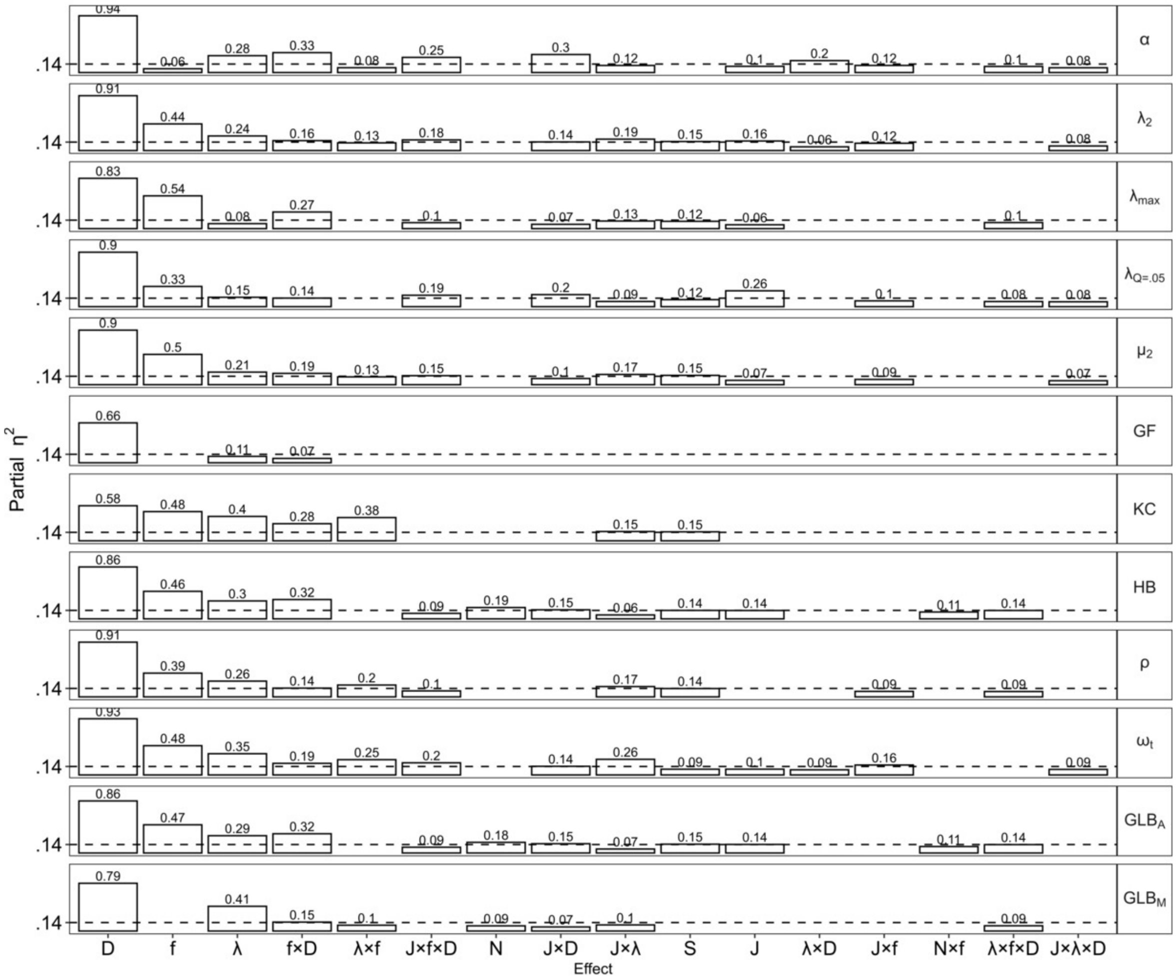

**Figure A5.** Effect sizes of particular factors and nontrivial interactions per coefficient for multidimensional conditions with non-normal distribution and continuous and discrete scale.