Monte Carlo simulation studies of reliability in psychometrics: A methodological review

Josip Novak*, Blaž Rebernjak

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

Abstract

Monte Carlo simulation studies are widely used in reliability research. This study reviewed 85 published Monte Carlo simulation studies investigating reliability. The review focused on the prevalence of particular reliability estimation methods and estimators, as well as adherence to previous recommendations for the Monte Carlo simulation method. It appears researchers do not fully adhere to these recommendations. Most of the reviewed studies have limitations in at least one of the following: reporting on the data generation procedure, selection of the number of replications, selection of conditions, benchmark utilization, and performance evaluation. Findings also suggest internal consistency in general and coefficient α are the most prevalent. Conversely, some reliability estimation methods and estimators that can be useful under many empirical conditions appear to be mostly overlooked. In the case of internal consistency, these are relatively obscure forms of α, λ₂, λ₄, μ-series, Kristof’s coefficient, Feldt-Gilmer coefficient, maximal reliability, greatest lower bound to reliability, ω family, structural equation modeling-based coefficients, and internal consistency confidence intervals. Overlooked reliability estimation methods are parallel forms, test-retest, multilevel reliability, latent class-based reliability, reliability of an individual, and Bayesian reliability. Suggestions for future research have been offered.

Keywords: Monte Carlo simulation, psychometrics, reliability, simulation study design, systematized review

*Corresponding author

Email addresses: josip.novak92@gmail.com (Josip Novak), brebernj@ffzg.hr (Blaž Rebernjak)
ORCID iDs:  (Josip Novak),  (Blaž Rebernjak)
1. Introduction

Monte Carlo simulation studies (MCSS) are computer experiments that involve a finite process of repeated generation of pseudo-random data and performance evaluation of target estimators.¹ Estimates provided by different estimators are dependent variables (DVs) under specified conditions in a factorial design. While the theoretical performance of estimators can be obtained analytically, MCSS are used when an analytical approach is not feasible, to investigate estimator robustness, estimators with unknown distribution, asymptotic estimator performance with small sample sizes, the degree up to which an estimator underperforms with violated assumptions, the simultaneous influence of several factors on estimators, or the interaction of assumption violation with other problems (Bandalos & Leite, 2013; Harwell et al., 1996).

Many different authors have provided tutorials and recommendations for conducting MCSS. Burton et al. (2006) and Morris et al. (2019) provided the recommendations in statistics, Bandalos and Leite (2013), Boomsma (2013), Paxton et al. (2001), and Skrondal (2000) for structural equation modeling (SEM), McNeish et al. (2018) for social sciences in general, and Harwell et al. (1996) and Feinberg and Rubright (2016) in psychometrics. They emphasize the methodological rigor in specific MCSS features that fall into the following categories: data generation procedure, the decision on the number of replications, selection of conditions in the design, benchmark² specification, and use of performance measures³ (see Bandalos & Leite, 2013; Boomsma, 2013; Burton et al., 2006; Feinberg & Rubright, 2016; Harwell et al., 1996; McNeish et al., 2018; Morris et al., 2019; Paxton et al., 2001; Skrondal, 2000). Reviews of the method are conducted on samples of published MCSS to evaluate up to what degree the researchers adhere to the recommendations.

1.1. Previous reviews of Monte Carlo simulation studies

So far, reviews of the MCSS method have been typically conducted in statistics. For instance, Hoaglin and Andrews (1975) observed a few limitations: not reporting which computer was used, using inaccurate software for data generation, and using defective pseudo-random number generators (RNG). Hauck and Anderson (1984) updated Hoaglin and Andrews (1975). They found the reporting on the RNG had improved, but researchers rarely justified the number of replications and rarely used statistical analysis. Burton et al. (2006) observed that most studies did not describe the data generation procedure to be replicable and that number of replications was typically not justified. Koehler et al. (2009) found a certain proportion of studies ignored Monte Carlo error (MCE)⁴ and did not report on using replications, while most studies did not justify the number.

Harwell et al. (1996) reviewed a small sample of MCSS in psychometrics and observed limitations such as the focus on relatively short tests, limited range of simulee sample size, conditions not corresponding to empirical, not providing supporting documentation, not implementing replications, and conducting oversimplified data analyses. Feinberg and Rubright (2016) also reviewed a small sample of MCSS in psychometrics. They found

---

¹Estimator is a function for calculating a quantity of interest based on the observed values.
²In simulation studies, benchmark represents the true value. Estimator values under specific conditions are compared against the benchmark to examine how dependable they are in estimating the true value.
³Performance measures should be discerned from the benchmark. While the benchmark represents the true value, performance measures are quantities that represents the estimator dependability in estimating the true value. Most common example of a performance measure is bias. Performance measures can also be DVs.
⁴MCE represents variability of results over multiple replications (Koehler et al., 2009).
researchers typically used maximally up to 500 replications per condition, did not justify the number of replications, or did not use replications at all. In a relatively recent review (Harwell et al., 2018), studies were sampled from both statistical and psychometric journals to update Hauck and Anderson (1984). The findings indicated the use of MCSS has increased over time, but the recommendations have not been adhered to.

1.2. Reliability

Psychometric theory yielded different definitions of reliability, which is generally the proportion of true score variance in the observed score variance. There are three main theoretical frameworks in psychometrics: classical test theory (CTT), generalizability theory (GT), and item response theory (IRT). CTT, also called true score theory, typically defines reliability as the ratio of the true to the observed score variance. Because multiple sources of unreliability exist, GT was presented as it enables the specification of the universe of conditions over which the results can be generalized (Cronbach et al., 1963). Based on criticisms of CTT, IRT was developed to address the fact that reliability is dependent on the latent parameter, often called ability, and to enable the calculation of marginal reliability based on the latent parameter (e.g., Hambleton & Van Der Linden, 1982). Reliability can also be viewed through factor-analytic (FA) framework, which is related to both CTT and IRT. For instance, when the congeneric model holds, true score variance is equal to the latent variable. Moreover, FA model for categorical variables can be reparametrized as an IRT model. There are numerous reliability estimators within these frameworks.

In the context of reliability, the largest number of reliability estimation methods and estimators have been presented within CTT. Guttman (1945) identified three sources of unreliability: individuals, items, and trials. There are three main CTT reliability estimation methods (Lord & Novick, 1967): test-retest, parallel forms, and internal analysis of item variances and covariances, which includes internal consistency and split-half reliability. Internal consistency coefficients, for instance, $\lambda_3$ (Guttman, 1945), widely but wrongly referred to as Cronbach’s $\alpha$, which represents the expected split-half value when the number of items is even (Cronbach, 1951), have become widely used since split-half coefficients depend on the approach to test division (Lord & Novick, 1967). Internal consistency can also be estimated using FA (e.g., McDonald, 1999). Some methods address the existence of individual differences. Reliability of an individual pertains to intra-individual variation between trials (see Fiske & Rice, 1955). Group differences can also be present, and variance occasionally exists at multiple levels. In such cases, it is justifiable to estimate reliability for each level separately (see Geldhof et al., 2014).

Among the available reliability estimation methods, internal consistency coefficients have been most commonly used in practical situations (e.g., Hogan et al., 2000). The most popular internal consistency coefficient is $\alpha$, whose indiscriminate use has caused discussions about its usefulness and suggestions of alternatives for particular conditions.

---

5 Reliability estimator represents an approach to obtaining the population reliability estimate as defined within a particular theoretical framework for particular reliability estimation method.

6 It should be noted the term internal consistency has not been clearly defined in the literature (e.g., S.-J. Cho et al., 2015; McDonald, 1999). Traditionally, internal consistency represents the extent to which the items of a test measure the attribute consistently—in particular having positive and “high” correlations (McDonald, 1999). Throughout the text, the term internal consistency is used following the traditional definition. However, a more appropriate term is required and there are suggested terms in the reviewed studies that might become more widespread in the future, such as single-test reliability (Pfadt et al., 2022) or single-administration reliability (Van Der Ark et al., 2011).
Various factors influence reliability estimates regardless of the theoretical framework. The selection of the appropriate estimator depends on empirical conditions. Reliability estimator performance is influenced by factors such as sample size, result distribution, missing data, presence of outliers, or population reliability value. Some factors stem from psychometric theory, such as item difficulty, item discrimination, inter-item covariation, or test length (see e.g., Lord & Novick, 1967), which implies the use of composite variables that consist of items. Composite variables have an underlying latent structure, which includes measurement model type, number of dimensions, and error correlatedness (e.g., Savalei & Reise, 2019a). There is evidence at least some of these factors are overlooked in practical situations as researchers rarely check the latent structure before reliability estimation (e.g., E. Cho, 2016).

1.3. Present study

MCSS is among the main approaches to investigating estimator performance under particular conditions. The standards for the conduct of MCSS in the field of statistics have been adopted to address the issue of reproducibility, as elaborated in previous reviews (e.g., Harwell et al., 2018). MCSS method in reliability significantly overlaps with other fields. Previous findings suggest the recommendations about the MCSS method have mostly not been implemented. This might be the case in the field of reliability as well, especially given that previous reviews in other fields suggest nonexistent to weak improvement over time (e.g., Harwell et al., 2018; Hauck & Anderson, 1984; Hoaglin & Andrews, 1975). Findings in psychometrics were similar to findings in other fields (Feinberg & Rubright, 2016; Harwell et al., 1996). However, they were based on small samples of studies, a limited range of evaluated features, and lacked a theoretical focus. Published studies with limitations in the method may occasionally be used as exemplary, which perpetuates these limitations, potentially beyond the field of reliability. Also, limitations in the method may result in inconsistent findings. A methodological review that would describe the practice up to date and facilitate the application of the method in future research is therefore needed.

Furthermore, a review of MCSS will indicate which methods, estimators, and factors have been overlooked and suggest future research in the field of reliability specifically. Substantive developments in reliability theory yielded numerous options for reliability estimation, but researchers have mostly used a limited range of these options (e.g., Hogan et al., 2000). E. Cho (2022) confirmed different estimators are appropriate in different conditions and it is known numerous factors affect reliability estimates by making them biased and imprecise. MCSS findings are used in estimator selection for particular conditions, but researchers are without guidance if particular factors that affect reliability are not sufficiently researched. However, it is not known which reliability estimation methods, estimators and factors lack attention in MCSS due to a lack of reviews. This lack of attention possibly influences practical reliability estimation and contributes to the use of a limited number of methods and estimators, misuse of particular estimators, and bad practices in reliability estimation such as overlooking the influence of particular factors. Therefore, this review addresses several gaps in the literature due to the general lack of methodological reviews of MCSS in psychometrics, in particular regarding reliability.
The criteria for study evaluation were selected to reflect the shared recommendations from different sources and cover key features in MCSS method categories. They also address the aforementioned gaps in reliability. To learn about relevant MCSS features outside the scope of this study, there are useful sources for the theoretical rationale, problem formulation (Bandalos & Leite, 2013; Boomsma, 2013; Harwell et al., 1996; McNeish et al., 2018; Morris et al., 2019; Paxton et al., 2001; Skrondal, 2000), and data management (Bandalos & Leite, 2013; Paxton et al., 2001). Features in the focus of this study are displayed in Table 1.

Therefore, this study aims to contribute with the following aims: (i) to describe the prevalence of particular reliability estimation methods and estimators, and (ii) to describe the prevalence of particular method features in published MCSS. The results are going to be compared to previous reviews. This review follows the protocol in Petticrew and Roberts (2008).

It is expected that CTT methods are the most prevalent in published MCSS, in particular internal consistency and $\alpha$. Due to being overlooked in practical reliability estimation, it is expected factors related to the latent structure are the least prevalent in published MCSS. Based on previous findings, it is also expected that adherence to the recommendations about the MCSS method in the field of reliability would echo previous reviews. Due to substantive development in reliability theory, heterogeneity over time is also expected.

2. Methods

2.1. Search strategy

Published studies employing MCSS to investigate reliability were searched in Google Scholar, Scopus, and Web of Science Core Collection (WoS) using combinations of terms “reliability”, “Monte Carlo”, and “simulation study” for each of these databases. These combinations of terms were selected to strike the balance between sensitivity and specificity (Petticrew & Roberts, 2008). Each out of the first 1000 available results in Google Scholar, the first 2000 available results in Scopus, and all of the results in WoS (1628, 3997, and 2442, for each combination of search terms respectively) were screened. Therefore, 17 067 studies were screened overall. The initial search was conducted from April to September 2021 and the final search was conducted in February 2022 to cross-validate the initial search using a more exhaustive procedure and increase the sample size. No specialized software was used for the search. Compared to prior reviews of MCSS, this study was not limited to particular journals but was limited to publications up to the year 2021 inclusive. These limitations were relaxed since the target population of studies is narrowly defined. Unpublished studies, conference papers, theses, and doctoral dissertations were not in the focus of the review. Studies were screened based on their title and abstract. After the screening, 208 studies were selected and 96 were left after duplicate removal. No automated tools were used in the selection.

2.2. Inclusion and eligibility

The eligibility criteria were the following: (i) the study is published in a peer-reviewed journal; (ii) the language of the study is English, and (iii) the study employed a simulation to investigate reliability estimators.

Provided protocol pertains to the final search.
Table 1: Study characteristics of identified relevant studies

<table>
<thead>
<tr>
<th>Features per category</th>
<th>Elaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimator selection:</strong> (i) Every investigated estimator; (ii) Theoretical framework.</td>
<td>These features have field-specific significance. Their selection is guided by research aims and theoretical framework.</td>
</tr>
<tr>
<td><strong>Data generation procedure:</strong> (i) Description of the data generation procedure; (ii) The model used for the data generation (e.g., FA or IRT); (iii) Software used for the data generation; (iv) Data validity check; (v) Code or simulation run output availability.</td>
<td>These are recommended features that ensure replicability and internal validity if conducted and reported (see Bandalos &amp; Leite, 2013; Boomsma, 2013; Burton et al., 2006; Feinberg &amp; Rubright, 2016; Harwell et al., 1996; McNeish et al., 2018; Morris et al., 2019; Paxton et al., 2001; Skrondal, 2000).</td>
</tr>
<tr>
<td><strong>Selection of the number of replications:</strong> (i) Number of replications; (ii) Justification for the number.</td>
<td>These features are comparable to statistical power. They reduce MCE and further increase internal validity if reported and properly applied (Bandalos &amp; Leite, 2013; Boomsma, 2013; Burton et al., 2006; Feinberg &amp; Rubright, 2016; Harwell et al., 1996; Morris et al., 2019; Skrondal, 2000, see).</td>
</tr>
<tr>
<td><strong>Selection of conditions:</strong> (i) Factors with levels; (ii) Justification for the selection of conditions.</td>
<td>These features are directly related to the external validity, which is improved with a larger number of conditions. MCSS are recommendably approached as a conventional experiment in terms of methodological rigor in the study design (see Bandalos &amp; Leite, 2013; Boomsma, 2013; Burton et al., 2006; Feinberg &amp; Rubright, 2016; Harwell et al., 1996; McNeish et al., 2018; Morris et al., 2019; Paxton et al., 2001; Skrondal, 2000). These features also have field-specific significance.</td>
</tr>
<tr>
<td><strong>Benchmark specification:</strong> (i) Whether the benchmark was specified; (ii) Whether the benchmark included multiple levels; (iii) Whether the benchmark was based on empirical data or finite simulated population.</td>
<td>These features facilitate performance evaluation as the true value is known. They also increase the external validity if specified according to empirical conditions (see Boomsma, 2013; Burton et al., 2006; Feinberg &amp; Rubright, 2016; Morris et al., 2019; Paxton et al., 2001).</td>
</tr>
<tr>
<td><strong>Performance evaluation:</strong> (i) Descriptive statistics used for performance evaluation; (ii) Inferential statistics used for performance evaluation; (iii) Performance measures; (iv) Whether visualization was used.</td>
<td>These features represent numerous performance evaluation methods whose selection depends on research aims and estimator(s) (see Bandalos &amp; Leite, 2013; Boomsma, 2013; Burton et al., 2006; Feinberg &amp; Rubright, 2016; Harwell et al., 1996; McNeish et al., 2018; Morris et al., 2019; Paxton et al., 2001; Skrondal, 2000).</td>
</tr>
</tbody>
</table>
Studies were not considered eligible if they focused on reliability generalization, reliability of classification, reliability of profiles, criterion-referenced reliability, rater agreement\(^8\), or if they investigated reliability of a specific scale. If it was not certain whether the studies were eligible, the method section was checked to see if it includes simulation, which was necessary only on a few occasions. Among the 96 studies, 40 were excluded. The list of excluded studies with reasons for exclusion is available in Appendix B. The initial sample size after the whole procedure was \(n = 56\). Therefore, studies were additionally searched by checking reference lists in the included studies, and \(n = 29\) additional studies were included based on this additional search. Studies sampled from the databases are commonly cited in many other studies from the sample. This suggests the sample is sufficiently representative of the relatively small target population. There are no duplicate publications in the sample of included studies. Because the target population are MCSS published in journals and one of the aims is describe the prevalence of their method features, no studies were excluded due to potential invalidity. Since the study sample is based on simulated data or simulees,\(^9\) the advantages are that potential bias stemming from the sampling method, any other bias stemming from participant characteristics in experimental designs, risk of bias in outcomes, and the potential need for the estimation of results from other information are eliminated. For these reasons and since no quantitative synthesis of estimator values was planned, no study was removed due to the potential risk of bias.

### 2.3. Coding

Coded features were those outlined in Table 1. The review and coding were limited to the simulation part in studies in which simulation was conducted jointly with the investigation of estimators using empirical data. If multiple simulations were conducted in a single article, coding for each simulation respectively was separated with “/”. In such cases, if data generation procedure, number of replications, or performance measures were identical for multiple simulations, single values were coded in corresponding columns that apply for every simulation and only coded conditions were separated with “/”. There was no need to contact the study authors for additional information. The coding was done by the first author and accuracy was checked two times after the initial coding with a time lag. This resulted in a correction of a few minor errors. The risk of subjective bias is minimized since the large majority of features were coded as reported. Exceptions are justifications for the number of replications and the justification for the selection of conditions. These were coded as binary variables and scored positively if at least partial justification was provided to minimize potential bias due to subjectivity. Only the observed information was coded and no assumptions about missing information were made. No software was used to extract the data. However, there is a certain risk of publication bias.

The protocol of the whole procedure is available in Appendix A. The overview of the included MCSS with coded characteristics in chronological order and corresponding references is provided in Appendix B. Footnotes for Tables 2 and 4 are provided in Appendix C.

### 2.4. Analysis

Analysis was conducted on feature frequencies and proportions. Features were also analyzed over time to examine the potential heterogeneity in that regard. The Cochrane-

---

\(^8\)Despite raters are among the typical measurement facets within GT, MCSS incorporating rater as one of the factors in the design were not considered eligible for inclusion in the sample because rater agreement is outside the focus of this study.

\(^9\)Simulee represents a simulated examinee (Feinberg & Rubright, 2016).
Armitage test was used to analyze trends in proportions with Cramer’s V as effect size (ES). The analysis was conducted using R (R Core Team, 2022) and package rstatix (Kassambara, 2021). The visualization was done using package ggplot2 (Wickham et al., 2022) and patchwork (Pedersen, 2020). Finally, the data were also analyzed narratively to indicate examples of potentially useful estimators for future research, comment on some observed practices, and inspect justifications for the selection of the number of replications and the selection of conditions.

3. Results and discussion

Table 2 provides a general description of the reviewed studies, including feature frequencies and percentages. The number of studies per year is displayed in Table 3.

<table>
<thead>
<tr>
<th>General feature</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of simulations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>81.2</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>11.8</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5.9</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>No. of estimators$^1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>24.1</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>25.3</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>14.5</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>12.1</td>
</tr>
<tr>
<td>$\geq$ 5</td>
<td>20</td>
<td>24.1</td>
</tr>
<tr>
<td>Equal no. of estimators</td>
<td>14</td>
<td>87.5</td>
</tr>
<tr>
<td>Different no. of estimators</td>
<td>2</td>
<td>12.5</td>
</tr>
<tr>
<td>No. of factors$^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>2.7</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>11.7</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>13.5</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>27.0</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>27.0</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>12.6</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>5.4</td>
</tr>
<tr>
<td>Theory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTT</td>
<td>74</td>
<td>87.1</td>
</tr>
<tr>
<td>IRT</td>
<td>2</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Continued on next page
Monte Carlo simulation studies of reliability in psychometrics

### Table 2: Frequencies and percentages of general study features (Continued)

<table>
<thead>
<tr>
<th>General feature</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTT &amp; IRT</td>
<td>7</td>
<td>8.2</td>
</tr>
<tr>
<td>GT</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Described procedure</td>
<td>82</td>
<td>96.5</td>
</tr>
<tr>
<td>Reported software</td>
<td>44</td>
<td>51.8</td>
</tr>
<tr>
<td>Checked data validity</td>
<td>19</td>
<td>22.4</td>
</tr>
<tr>
<td>Provided code or output</td>
<td>5</td>
<td>5.9</td>
</tr>
<tr>
<td>Reported no. of replications</td>
<td>79</td>
<td>92.9</td>
</tr>
<tr>
<td>Justified no. of replications</td>
<td>8</td>
<td>9.4</td>
</tr>
<tr>
<td>Justified conditions</td>
<td>36</td>
<td>42.4</td>
</tr>
<tr>
<td>Unidimensional</td>
<td>61</td>
<td>71.8</td>
</tr>
<tr>
<td>Parallel</td>
<td>40</td>
<td>47.1</td>
</tr>
<tr>
<td>Uncorrelated errors</td>
<td>65</td>
<td>76.5</td>
</tr>
<tr>
<td>Continuous scale</td>
<td>45</td>
<td>52.9</td>
</tr>
<tr>
<td>Normal distribution</td>
<td>35</td>
<td>41.2</td>
</tr>
<tr>
<td>Homogeneous population</td>
<td>77</td>
<td>90.6</td>
</tr>
<tr>
<td>Purely theoretical conditions</td>
<td>9</td>
<td>10.6</td>
</tr>
<tr>
<td>Specified benchmark</td>
<td>69</td>
<td>81.2</td>
</tr>
<tr>
<td>Used descriptive statistics³</td>
<td>49</td>
<td>57.7</td>
</tr>
<tr>
<td>Used inferential statistics</td>
<td>26</td>
<td>30.6</td>
</tr>
<tr>
<td>Used performance measures</td>
<td>56</td>
<td>65.9</td>
</tr>
<tr>
<td>Used visualization</td>
<td>41</td>
<td>48.2</td>
</tr>
</tbody>
</table>

**Notes:** n = no. of studies, Unidimensional = unidimensionality in every investigated condition, Parallel = parallel measurement model in every investigated condition, Uncorrelated errors = uncorrelated errors in every investigated condition, Continuous scale = continuous scale in every condition, Normal distribution = normal distribution in every condition, Purely theoretical = unidimensional parallel measurement model with uncorrelated errors, continuous scale, normal distribution, and homogeneous population in every condition. For footnotes see Appendix C.

### Table 3: Number of studies per year

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1966</td>
<td>1</td>
</tr>
<tr>
<td>1969</td>
<td>1</td>
</tr>
<tr>
<td>1973</td>
<td>1</td>
</tr>
<tr>
<td>1975</td>
<td>1</td>
</tr>
<tr>
<td>1976</td>
<td>1</td>
</tr>
<tr>
<td>1977</td>
<td>2</td>
</tr>
<tr>
<td>1978</td>
<td>1</td>
</tr>
<tr>
<td>1980</td>
<td>1</td>
</tr>
<tr>
<td>1986</td>
<td>1</td>
</tr>
<tr>
<td>1987</td>
<td>1</td>
</tr>
<tr>
<td>1990</td>
<td>1</td>
</tr>
<tr>
<td>1992</td>
<td>1</td>
</tr>
<tr>
<td>1993</td>
<td>1</td>
</tr>
<tr>
<td>1995</td>
<td>1</td>
</tr>
<tr>
<td>1996</td>
<td>1</td>
</tr>
<tr>
<td>1997</td>
<td>3</td>
</tr>
<tr>
<td>1999</td>
<td>3</td>
</tr>
<tr>
<td>2000</td>
<td>2</td>
</tr>
<tr>
<td>2003</td>
<td>1</td>
</tr>
<tr>
<td>2004</td>
<td>2</td>
</tr>
<tr>
<td>2006</td>
<td>4</td>
</tr>
<tr>
<td>2007</td>
<td>4</td>
</tr>
<tr>
<td>2008</td>
<td>3</td>
</tr>
<tr>
<td>2009</td>
<td>1</td>
</tr>
<tr>
<td>2010</td>
<td>4</td>
</tr>
<tr>
<td>2011</td>
<td>2</td>
</tr>
<tr>
<td>2012</td>
<td>8</td>
</tr>
<tr>
<td>2013</td>
<td>4</td>
</tr>
<tr>
<td>2014</td>
<td>3</td>
</tr>
<tr>
<td>2015</td>
<td>3</td>
</tr>
<tr>
<td>2016</td>
<td>6</td>
</tr>
<tr>
<td>2017</td>
<td>1</td>
</tr>
<tr>
<td>2018</td>
<td>4</td>
</tr>
<tr>
<td>2019</td>
<td>2</td>
</tr>
<tr>
<td>2020</td>
<td>3</td>
</tr>
<tr>
<td>2021</td>
<td>5</td>
</tr>
</tbody>
</table>
3.1. General features of reviewed studies

Researchers mostly used one simulation per study. They commonly investigated one, compared two estimators, or compared five or more estimators. If multiple simulations were conducted, they mostly investigated an equal number of estimators. Furthermore, researchers most often manipulated three or four factors in the design, followed by two and five, but never over six. The most prevalent theoretical framework was expectedly CTT, followed by the combination of CTT and IRT. IRT and GT appear to be less prevalent.

3.2. General data generation procedure features

A large majority of researchers described the data generation procedure, but they have typically not reported which software was used for the data generation. Reporting about or doing any form of the generated data validity check appears to be relatively uncommon, especially providing the code used for the data generation, or at least simulation run output. These results are mostly in line with previous reviews (e.g., Burton et al., 2006; Harwell et al., 2018; Hauck & Anderson, 1984), except for the description of the data generation procedure, which appears to be more detailed in the field of reliability.

3.3. General reporting on the number of replications and justification for the number

Researchers mostly reported the number of replications, but the number was typically not justified, which is in line with previous findings (Burton et al., 2006; Harwell et al., 2018; Hauck & Anderson, 1984; Koehler et al., 2009; Morris et al., 2019). However, provided justifications are not necessarily in line with guidelines.

3.4. General features in terms of factors affecting reliability

The selection of conditions was not justified in more than half studies, which echoes Harwell et al. (2018). Most of the studies investigated the estimators under the condition of unidimensionality, uncorrelated errors, and a homogeneous population. Furthermore, it also appears to be relatively common to use continuous scale, parallel measurement model, or normally distributed results in every condition. However, purely theoretical conditions (unidimensionality, parallel measurement model, uncorrelated errors, continuous scale, and normal distribution present simultaneously) are uncommon.

3.5. General features of benchmark specification and performance evaluation methods

Furthermore, most of the studies specified a benchmark, but in a certain proportion of studies, this specific advantage was not used.

Finally, performance measures were used in most of the studies to evaluate estimator performance, followed by descriptive statistics and inferential statistical techniques. It appears researchers in the field of reliability use the latter relatively often (cf. Harwell et al., 2018; Hauck & Anderson, 1984). Some form of data visualization used for performance evaluation was used in approximately half of the reviewed studies. The features are inspected more closely in Table 4.

<table>
<thead>
<tr>
<th>Table 4: Frequencies and percentages of specific study features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
</tr>
<tr>
<td>Estimator</td>
</tr>
</tbody>
</table>

Continued on next page
Table 4: Frequencies and percentages of specific study features (Continued)

<table>
<thead>
<tr>
<th>Feature</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal consistency</td>
<td>58</td>
<td>68.2</td>
</tr>
<tr>
<td>Variants of $\alpha$</td>
<td>43</td>
<td>50.6</td>
</tr>
<tr>
<td>$\alpha^1$</td>
<td>39</td>
<td>45.9</td>
</tr>
<tr>
<td>Stratified $\alpha$</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Standardized $\alpha$</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Obscure forms of $\alpha$</td>
<td>7</td>
<td>8.2</td>
</tr>
<tr>
<td>Kuder-Richardson $^2$</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>5</td>
<td>5.9</td>
</tr>
<tr>
<td>$\lambda_5$</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>$\lambda_6$</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>$\mu$-series</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>Kristof</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Feldt-Gilmer</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Raju</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Maximal reliability</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>$\omega$ family $^3$</td>
<td>15</td>
<td>17.7</td>
</tr>
<tr>
<td>$\omega_{\text{hierarchical}}$</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>$\omega_{\text{categorical}}$</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>Greatest lower bound (GLB)$^4$</td>
<td>6</td>
<td>7.1</td>
</tr>
<tr>
<td>SEM-based$^5$</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Split-half</td>
<td>10</td>
<td>11.8</td>
</tr>
<tr>
<td>Parallel forms</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>Test-retest</td>
<td>6</td>
<td>7.1</td>
</tr>
<tr>
<td>CIs</td>
<td>16</td>
<td>18.8</td>
</tr>
<tr>
<td>CIs of any variant of $\alpha$</td>
<td>13</td>
<td>15.3</td>
</tr>
<tr>
<td>CIs of any variant of $\omega$</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>IRT-based estimator</td>
<td>9</td>
<td>10.6</td>
</tr>
<tr>
<td>Generalizability coefficient</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Reliability of an individual</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Multilevel reliability</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Latent class-based reliability</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>Bayesian reliability</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Performance evaluation$^6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M)ANOVA</td>
<td>22</td>
<td>25.9</td>
</tr>
<tr>
<td>Regression analysis</td>
<td>4</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Continued on next page
Table 4: Frequencies and percentages of specific study features (Continued)

<table>
<thead>
<tr>
<th>Feature</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation/coefficient of determination</td>
<td>5</td>
<td>5.9</td>
</tr>
<tr>
<td>Bias</td>
<td>26</td>
<td>30.6</td>
</tr>
<tr>
<td>Relative bias</td>
<td>11</td>
<td>12.9</td>
</tr>
<tr>
<td>Absolute bias</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Precision (SD or variance)</td>
<td>22</td>
<td>25.9</td>
</tr>
<tr>
<td>(R)MSE</td>
<td>18</td>
<td>21.2</td>
</tr>
<tr>
<td>Coverage</td>
<td>18</td>
<td>21.2</td>
</tr>
<tr>
<td>Benchmark</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple levels</td>
<td>26</td>
<td>30.6</td>
</tr>
<tr>
<td>Finite simulated population</td>
<td>10</td>
<td>11.8</td>
</tr>
<tr>
<td>Based on empirical data</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Data generation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>13</td>
<td>15.3</td>
</tr>
<tr>
<td>SAS</td>
<td>9</td>
<td>10.6</td>
</tr>
<tr>
<td>Mplus</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>SPSS</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>EQS</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>MATLAB</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>FORTRAN</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>9.4</td>
</tr>
<tr>
<td>FA</td>
<td>21</td>
<td>24.7</td>
</tr>
<tr>
<td>IRT</td>
<td>12</td>
<td>14.1</td>
</tr>
<tr>
<td>Time series</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Other</td>
<td>48</td>
<td>55.3</td>
</tr>
<tr>
<td>Generation based on empirical data</td>
<td>9</td>
<td>10.6</td>
</tr>
<tr>
<td>Validity checked using model fit/convergence</td>
<td>9</td>
<td>10.6</td>
</tr>
<tr>
<td>Validity checked using different software</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Validity checked using cross-validation</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Other forms of validity check</td>
<td>6</td>
<td>7.1</td>
</tr>
<tr>
<td>Provided code</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Provided output</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>Replications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 99</td>
<td>11</td>
<td>12.6</td>
</tr>
<tr>
<td>100–999</td>
<td>31</td>
<td>35.6</td>
</tr>
<tr>
<td>1000–4999</td>
<td>30</td>
<td>34.5</td>
</tr>
<tr>
<td>5000–9999</td>
<td>2</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Continued on next page
Table 4: Frequencies and percentages of specific study features (Continued)

<table>
<thead>
<tr>
<th>Feature</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 10,000$</td>
<td>13</td>
<td>14.9</td>
</tr>
<tr>
<td>100</td>
<td>16</td>
<td>18.4</td>
</tr>
<tr>
<td>500</td>
<td>6</td>
<td>6.9</td>
</tr>
<tr>
<td>1000</td>
<td>23</td>
<td>26.4</td>
</tr>
<tr>
<td>2000</td>
<td>6</td>
<td>6.9</td>
</tr>
<tr>
<td>10,000</td>
<td>6</td>
<td>6.9</td>
</tr>
<tr>
<td>Multiple no. of replications</td>
<td>8</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Factors

<table>
<thead>
<tr>
<th>Feature</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulee sample size</td>
<td>45</td>
<td>52.9</td>
</tr>
<tr>
<td>No. of items</td>
<td>44</td>
<td>51.8</td>
</tr>
<tr>
<td>No. of scale points</td>
<td>24</td>
<td>28.2</td>
</tr>
<tr>
<td>Distribution shape (any)</td>
<td>23</td>
<td>27.1</td>
</tr>
<tr>
<td>Distribution (observed)</td>
<td>19</td>
<td>22.4</td>
</tr>
<tr>
<td>Distribution (true scores)</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Distribution (error scores)</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Item difficulty</td>
<td>5</td>
<td>5.9</td>
</tr>
<tr>
<td>Item discrimination</td>
<td>7</td>
<td>8.2</td>
</tr>
<tr>
<td>Inter-item covariation</td>
<td>9</td>
<td>10.6</td>
</tr>
<tr>
<td>Measurement model type/factor loadings</td>
<td>27</td>
<td>31.8</td>
</tr>
<tr>
<td>Error correlatedness</td>
<td>8</td>
<td>9.4</td>
</tr>
<tr>
<td>No. of dimensions</td>
<td>6</td>
<td>7.1</td>
</tr>
<tr>
<td>Hierarchical structure</td>
<td>6</td>
<td>7.1</td>
</tr>
<tr>
<td>Missingness</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>Outliers</td>
<td>4</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Notes: For footnotes see Appendix C.

3.6. Prevalence of particular estimators

Internal consistency coefficients and $\alpha$ are the most prevalent, as expected. Split-half reliability, test-retest, and parallel forms reliability are less prevalent compared to internal consistency. This is in line with previous findings concerning the typically used methods to examine reliability in research practice (e.g., Hogan et al., 2000). Many reliability estimation methods and estimators appear to be overlooked in general.

3.7. Specific features of the data generation procedure

Researchers mostly used R for data generation, which is in line with Morris et al. (2019), followed by SAS, Mplus, and SPSS. If the data were generated following a specific model, it was most often done using the FA model, followed by the IRT model. Data generation was mostly done using RNGs without following a specific model. Empirical datasets were rarely used as a basis for the data generation, which is in line with Burton et al. (2006). To check the generated data validity, the most common were model fit examination and/or convergence,
followed by the check using different software and some form of cross-validation for at least some conditions. In the rare cases of providing them, researchers appear to be more inclined towards providing the code \( (n = 4) \) than providing the simulation run output \( (n = 1) \).

3.8. Specific reporting on the number of replications and justification for the number

Furthermore, researchers mostly used up to 1000 replications, most often exactly 1000, followed by 100. Researchers also mostly used a single number of replications for every condition. Previous research yielded similar results. Out of 223 studies, Koehler et al. (2009) reviewed, 74 used 1000 replications, whereas only five used over 10 000 replications. Burton et al. (2006) and Robey and Barcikowski (1992) also discovered that most researchers used 1000 replications, while Feinberg and Rubright (2016) observed that researchers typically used up to 500 replications.

3.9. Prevalence of particular factors affecting reliability

Simulee sample size appears to be the most prevalent factor in the research design, followed by the number of items, measurement model type or any manipulation with factor loadings, number of scale points, and manipulation with distribution shape. Regarding the latter, it appears observed scores are more often transformed into a non-normal shape compared to the transformation of true scores or error scores specifically. On the other hand, inter-item covariation, error correlatedness, item discrimination, number of dimensions, hierarchical structure, item difficulty, the impact of outliers, and missing data on reliability estimators are less prevalent. Therefore, the expectation that factors related to the latent structure would be the least prevalent in published MCSS is partially confirmed.

3.10. Specific features of the benchmark specification

When a benchmark was specified, a certain proportion of studies used multiple levels of a benchmark for the same combination of factors. It appears it is less common to specify a benchmark based on a generated finite population or based on the properties of an empirical dataset.

3.11. Prevalence of specific performance evaluation methods

If inferential statistical techniques were used for performance analysis, forms of ANOVA were typically selected, which is in line with Harwell et al. (2018), followed by correlation and regression analysis. Bias appears to be the most prevalent among the performance measures, followed by precision, (R)MSE, and coverage. This is in line with Morris et al. (2019) and partially in line with Harwell et al. (2018).

3.12. Analysis of trends over time

Since the studies are unequally represented across the unevenly spaced years, the years were grouped as follows: (i) years up to 1999, (ii) years from 2000 to 2012, and (iii) years from 2013 to 2021. Therefore, the first group consists of \( n = 23 \) studies, the second group consists of \( n = 29 \) studies, and the third group consists of \( n = 33 \) studies. Such grouping was selected to make the groups relatively similar in size while acknowledging that MCSS are increasingly often conducted (Harwell et al., 2018). Specific features to be analyzed over time were selected to broadly cover every method category in Table 1. Features pertaining to reliability methods, estimators, and factors were limited to the five most frequent ones. Trends over time are displayed in Figure 1.
Figure 1: Analysis of trends over time
3.13. Trends over time in the prevalence of particular estimators

Figure 1 illustrates that the prevalence of the $\omega$ family displays an upward trend with medium ES ($\chi^2(1) = 8.51, p = 0.004, V = 0.316$). On the other hand, internal consistency in general ($\chi^2(1) = 0.116, p = 0.733, V = 0.037$) and various forms of $\alpha$ coefficient ($\chi^2(1) = 0.082, p = 0.775, V = 0.031$) do not display any significant upward trend with a non-trivial ES. Furthermore, $\alpha$ specifically ($\chi^2(1) = 0.492, p = 0.483, V = 0.076$), and CIs of internal consistency coefficients ($\chi^2(1) = 0.627, p = 0.428, V = 0.086$) do not display any significant upward trend, but the ES is small.

3.14. Trends over time in data generation features

It appears describing the procedure ($\chi^2(1) = 0.88, p = 0.348, V = 0.102$) and conducting a data validity check ($\chi^2(1) = 0.06, p = 0.804, V = 0.027$) do not display any significant upward trend with non-trivial ES. On the other hand, reporting the software used for the data generation ($\chi^2(1) = 5.69, p = 0.017, V = 0.259$) and providing code or output ($\chi^2(1) = 3.84, p = 0.050, V = 0.212$) appear to display a significant upward trend with medium and low to medium ES, respectively.

3.15. Trends over time in reporting on the number of replications and justification for the number

The number of replications was log-transformed to base 10 to make the distribution more symmetrical and reduce heteroscedasticity. One-way ANOVA was applied, which suggests the upward trend is significant with approximately medium ES ($F(1, 85) = 4.472, p = 0.037, \eta^2 = 0.05$). However, providing the justification for the number of replications appears to have been flat and does not appear to display a significant downward trend afterward, but the ES is small to medium ($\chi^2(1) = 2.30, p = 0.129, V = 0.163$).

3.16. Trends over time in the prevalence of particular factors affecting reliability

The prevalence of simulee sample size ($\chi^2(1) = 6.90, p = 0.009, V = 0.285$), number of items ($\chi^2(1) = 2.48, p = 0.115, V = 0.171$), and measurement model type/factor loadings ($\chi^2(1) = 1.96, p = 0.162, V = 0.152$) in the design appear to display a significant and non-significant upward trend with medium, low to medium, and low to medium ES, respectively. The number of scale points ($\chi^2(1) = 0.43, p = 0.514, V = 0.071$) and distribution ($\chi^2(1) = 0.68, p = 0.411, V = 0.090$) do not display any significant upward trend with non-trivial ES. Finally, providing at least partial justification for the selection of conditions ($\chi^2(1) = 5.74, p = 0.017, V = 0.300$) appears to display a significant upward trend with medium ES.

3.17. Trends over time in the benchmark specification

Benchmark specification displays a significant upward trend with medium ES ($\chi^2(1) = 11.66, p = 0.001, V = 0.370$). However, the prevalence of a benchmark with multiple levels ($\chi^2(1) = 1.42, p = 0.234, V = 0.129$), a benchmark based on empirical data ($\chi^2(1) = 1.45, p = 0.228, V = 0.131$), or benchmark based on finite simulated population ($\chi^2(1) = 1.40, p = 0.237, V = 0.128$) do not appear to display any significant upward trend and the ES is small.
3.18. Trends over time in performance evaluation

To evaluate performance, the use of descriptive statistics does not appear to display any significant upward trend with non-trivial ES ($\chi^2(1) = 0.00, p = 0.949, V = 0.007$), as well as the use of inferential statistics ($\chi^2(1) = 0.10, p = 0.756, V = 0.034$), which appears to be equally often used over time. The use of visualization appears to have increased and remained flat with a significant trend and medium ES ($\chi^2(1) = 4.88, p = 0.027, V = 0.240$). Finally, the use of performance measures appears to display a significant upward trend with medium ES ($\chi^2(1) = 4.46, p = 0.035, V = 0.229$).

3.19. Prevalence of particular estimators

The results (Table 2) indicate many existing methods and reliability estimators have a low prevalence in published MCSS, such as IRT-based estimators, coefficients of generalizability, split-half, parallel forms, test-retest, internal consistency CIs, multilevel reliability, latent class-based reliability, reliability of an individual, or Bayesian reliability. CTT framework, internal consistency, and $\alpha$ in particular were expectedly shown to be the most prevalent. However, many internal consistency estimators were shown to have a low prevalence, such as $\lambda_2, \lambda_4, \mu$-series, Kristof’s coefficient, Feldt-Gilmer coefficient, maximal reliability, $\omega$ family, GLB, and SEM-based coefficients. Some of the methods and estimators are outdated for some purposes, such as split-half coefficients which can be used to estimate reliability of speeded tests, but are generally outdated as internal consistency coefficients. Some other methods and estimators are less commonly used, such as test-retest and parallel forms reliability, so these are expectedly less prevalent. Still, reliability estimation practice would possibly improve if particular overlooked methods and estimators received more attention in future research, especially internal consistency which is the most prevalent.

Internal consistency coefficients are theoretically interrelated and the selection of coefficients to investigate can be guided by that fact. E. Cho (2016) argued reliability coefficients “should be understood as building blocks of a single system rather than as a collection of completely unrelated methods.” Furthermore, E. Cho (2022) reanalyzed a certain number of MCSS of internal consistency coefficients and concluded that there is no single, most accurate reliability estimator among the analyzed coefficients and that a variety of coefficients are under-researched. The results suggest (Table 1) that in approximately half the studies up to two estimators were investigated. It is difficult to find a plausible reason not to compare a larger number of estimators in the same study. Some of the positive examples are Zijlmans, van der Ark, et al. (2018) and Osburn (2000) who decided to investigate several coefficients. If a set of coefficients is investigated without mathematical justification, then the recommendation about the appropriateness of a certain coefficient is more limited, since more mathematically appropriate coefficients (in some cases theoretically greater lower bounds to reliability) have been potentially overlooked. Therefore, the selection of coefficients based on their expected inter-relationships probably results in a greater contribution to theoretical articles.

$\alpha$ is the most used coefficient in practice (e.g., Hogan et al., 2000) and results of this study suggest $\alpha$ is also the most investigated in MCSS. The conclusion in many reviewed studies (e.g., Hogan et al., 2000; Thompson et al., 2010; Zimmerman et al., 1993) is that $\alpha$ is often a biased reliability estimator of limited usefulness (Bentler, 2021; Green & Yang, 2009; Raykov & Marcoulides, 2019; Sijtsma, 2009; Sijtsma & Pfadt, 2021). The malpractice of giving priority to $\alpha$ can be explained by tradition and, ultimately, ease of calculation (e.g., Sijtsma, 2009), but it is also possibly facilitated by the excessive focus of MCSS on $\alpha$ and neglect of alternatives. The analysis of trends over time suggests the internal consistency
coefficients generally and α variants are equally often researched over time. Conversely, an insignificant upward trend with small ES was detected for α specifically and internal consistency CIs.

There are different occasionally useful but obscure forms of α observed in the reviewed studies. Examples are α corrected for error covariance (Komaroff, 1997), α robust to outliers (Christmann & Van Aelst, 2006), α based on polychoric correlation useful for ordinal data (Zumbo et al., 2007), α based on L-comoments useful for heavy-tailed distributions and small sample sizes (Headrick & Sheng, 2013), IRT-variant of α useful for mixed item formats and automated test assembly (Shu & Schwarz, 2014), α based on copula function useful for non-normal ordinal data (Bonanomi et al., 2015), and α robust to both outliers and missing data (Zhang & Yuan, 2016).

ω family is the most prevalent after α and its prevalence displays an upward trend with medium ES. It should be noted that ω represents a family of coefficients and the context should be taken into account to know which one is used (e.g., E. Cho, 2016, 2021). ω coefficients depend on the FA method. There are few examples of MCSS investigating FA methods and ω family (e.g., Fu et al., 2021; Garcia-Garzon et al., 2021; Zinbarg et al., 2006, 2007). However, the ω family, as well as some other FA-based coefficients, such as GLB (P. H. Jackson & Agunwamba, 1977), ordinal θ (Zumbo et al., 2007), generalized Kristof’s coefficient (Ten Berge & Sočan, 2004), generalized θ (Şimşek & Noyan, 2013), or SEM-based coefficients (e.g., Raykov, 1997; Raykov & Marcoulides, 2015) occasionally outperform coefficients based on item variances and covariances, but should, of course, not be used indiscriminately (e.g., E. Cho, 2022).

Point estimates provide only partial information about the population value, but the prevalence of internal consistency CIs is low in the reviewed studies. Zhang and Yuan (2016) also presented robust variants of ω, as well as α and ω CIs. Most CI research is related to α or ω, while CIs of other coefficients are typically not investigated. However, the insignificant upward trend of CI prevalence displays small ES.

3.20. Data generation

A large majority of reviewed studies described the data generation procedure (Table 2) and have done that approximately equally over time. However, only approximately half reported the software used for data generation, which is possibly an issue since differences in generated data between software might exist (Burton et al., 2006; Morris et al., 2019). The analysis of trends over time suggests reporting about the software used for the data generation displays a significant upward trend with medium ES (Figure 1).

It is also recommended to select the model for the data generation. Since reliability estimates are based on measurement models, FA and IRT models can be particularly useful in MCSS. Among specific models, data were most often generated in line with FA, followed by the IRT model. An advantage of the FA or IRT model, is they are more intuitively understandable to researchers who are potentially going to use the findings to select an appropriate estimator. Using RNG and transforming the data can occasionally be required if factors except those related to the latent structure are also included, but generating data according to FA or IRT model (see Fife et al., 2012; Shu & Schwarz, 2014, for more details) is apt if the conditions represent a single time point. Time series models are useful for investigating coefficients of stability and reliability of an individual (see Du & Wang, 2018; Wang & Grimm, 2012, for more details). GT model, while not popular and often not feasible in practice, can be useful for data generation in a longitudinal context (see e.g., Laenen et al., 2008). Generating data according to empirical datasets result in an increased external
validity. Table 2 suggests such practice was found in few examples (e.g., Kim & Feldt, 2010; Ten Berge & Sočan, 2004).

Supporting documentation should be provided together with the evidence of data validity. Table 2 suggests providing the output from a computer run and reporting on the data validity check is uncommon, which Hoaglin and Andrews (1975) and its most recent update (Harwell et al., 2018). Data validity check also does not appear to improve over time (Figure 1). It is recommended to check whether the generated data are in line with specified parameters, and generate anew if not, as done in some studies (e.g., Barchard & Hakstian, 1997a, 1997b; Padilla & Divers, 2013). This is also applicable to the benchmark (e.g., Padilla et al., 2012). Checking model fit and/or convergence is required if, for instance, the FA model was used for the generation, as done by Zinbarg et al. (2006, 2007) and Yang and Xia (2019). Cross-validation using different software is possibly the safest approach to avoid misleading conclusions (e.g., Greer et al., 2006). Furthermore, there are a few positive examples with the full code available (e.g., S.-J. Cho et al., 2019). Another option is providing the output from a computer run (e.g., Edwards et al., 2021). The analysis of trends over time suggests that making the code or simulation run output available displays an upward trend (Figure 1).

3.21. Number of replications

Minimizing MCE requires a larger number of replications than commonly used in MCSS (Koehler et al., 2009). The magnitude of MCE and the required number of replications is a function of both the experimental design complexity and the target estimator. If technical prerequisites are fulfilled and sufficient time is available, there is no upper limit on the number of replications. For instance, Sheng and Sheng (2012) used 100,000 replications. However, such a large number is not always required. Skrondal (2000) observed that with simplistic designs a smaller number of replications can be sufficient. To avoid potentially misleading conclusions, it is generally sensible to aim as high as feasible, especially with a relatively complex research design. Due to the state of current knowledge in the field of reliability, simplistic designs are probably not very useful in many cases and the fact that the median number of replications has increased over time is in line with that (Figure 1).

In the context of reliability estimation, even a minor improvement can be considered relevant, which makes MCSS results in the field especially sensitive to the number of replications. Thus, the design of MCSS, upon whose findings researchers often rely, should be conducted with the awareness that MCE can confound the results and potentially shadow or overemphasize the observed differences in the estimator performance. Therefore, some reviewed studies potentially used an insufficient number of replications (Koehler et al., 2009).

However, it is currently impossible to exactly determine the required minimum number of replications for specific estimators. Robey and Barcikowski (1992) presented a procedure for objective determination of the number of replications required to detect departures from robustness. Such procedure is yet to be presented for most reliability estimators, which is required since the analysis of trends over time suggests that justifying the number of replications displays a slightly decreasing trend with small to medium ES (Figure 1). The following examples of justification for the number of replications were observed: to adequately account for the variability inherent in the simulation process (Bost, 1995), to ensure that the slightest departure from nominal levels would be detected (Barchard & Hakstian, 1997a, 1997b), due to time length required to calculate $\lambda_4$ (Thompson et al., 2010), based on the recommendation of Robey and Barcikowski (1992) and Romano et al. (2010, 2011), as an arbitrary balance between generating a large enough number of replications
to obtain appropriate precision of our estimates and the time required to analyze, and because that number was used in previous studies (Geldhof et al., 2014), and based on a previous study and literature (Gurdil Ege & Demir, 2020; Harwell et al., 1996). Most of these justifications are more or less arbitrary and none of them are objective, except for the one by Romano et al. (2010, 2011), but the justification referring to Robey and Barcikowski (1992) is limited to internal consistency CIs. Moreover, such procedure would also depend on the performance aspect in the focus, since a larger number of replications would probably be required to investigate estimator precision than to investigate bias.

### 3.22. Selection of conditions

It is recommended to justify the selection of conditions in MCSS. Considering the abundance of psychometric literature, not justifying the selection of factors and their levels is as fallacious as arbitrarily hypothesizing relationships in a theoretical model without regard for previous findings. Some positive examples of clear reporting on experimental design and justifying the selection of factor levels are found among the reviewed studies (e.g., S.-J. Cho et al., 2019; Kelley & Pornprasertmanit, 2016). The trend of justifying the selection of conditions appears to be upward with medium ES (Figure 1).

Many different justifications for the selection of conditions were observed. Some examples are: based on previous literature generally (e.g., Fu et al., 2021; Gurdil Ege & Demir, 2020; Y. Liu & Zumbo, 2007; Yang & Xia, 2019), factor levels were justified to be comparable to previous research (e.g., Padilla & Divers, 2016) based on previous studies (e.g., Trinchera et al., 2018), to examine a broader range of conditions compared to previous research (e.g., Zinbarg et al., 2007), to consider a range of conditions that can be observed or expected in real situations (e.g., Cheng et al., 2012; Edwards et al., 2021; Trizano-Hermosilla et al., 2021), or factor levels justified based on both previous studies and occurrence in empirical data (e.g., Turner et al., 2017). It is desirable to design research based on previous findings and specify conditions based on typical empirical conditions or literature. However, in none of the reviewed studies was the justification also based on the expected interaction between factor levels, which could result in useful findings, especially with external validity in mind.

This raises the issue of design complexity, which is inextricable from computational efficiency. If researchers decide to use a complex experimental design to increase external validity and possibly computational efficiency, an incomplete factorial design can be useful (see Harwell et al., 2017; Skrondal, 2000, for more information). In the case of reliability, this should be done with extreme caution not to omit some potentially useful insights about the interactions. On the other hand, incomplete factorial design may be a justifiable choice if a study follows up on previous research by expanding the varied conditions in which part of the design and reliability estimators are identical, as well as the software and/or seed used for data generation.

However, follow-up on previous research and replication studies are uncommon in the field of reliability. For instance, Jenkins and Taber (1977) followed up on Lissitz and Green (1975), Zinbarg et al. (2007) followed up on Zinbarg et al. (2006), Y. Liu et al. (2010) followed up on Zumbo et al. (2007), Padilla et al. (2012) followed up on Maydeu-Olivares et al. (2007), Edwards et al. (2021) extended the work of Green and Yang (2009), whereas Trizano-Hermosilla and Alvarado (2016) used identical distributions as Sheng and Sheng (2012). Moreover, none of the reviewed studies was an exact replication of some previous studies using different software, which could be encouraged as a plausible justification for conducting MCSS.
Reliability is influenced by various factors that interact at inducing positive or negative\textsuperscript{10} bias in reliability estimates. A lot of reviewed MCSS have overlooked interactions between different sources of bias. When exploring interactions using ANOVA, conclusions are at least partially dependent on selected factor levels and the number of factor levels so they should be carefully specified. Skrondal (2000) observed that third- or higher-order interactions are occasionally discarded to achieve parsimony and focus on the essentials of the investigated relationships. He added that at some point higher-order interactions can become negligible, as well as that interpretation of higher-order interactions is strenuous. However, in the field of reliability, when the performance of well-known coefficients is investigated up to a certain point with certain combinations of factors that are typical in practice, probing for higher-order interactions can be a potentially useful follow-up. It can be especially useful since most of the studies included up to four factors in the design (Table 2). For instance, it has been overlooked which factors can compensate for levels of other factors that make reliability estimators more biased, and possibly imprecise.

Furthermore, some factors are more prevalent in the design and others are overlooked. Table 4 shows simulee sample size is the most prevalent, followed by the number of items, number of scale points, distribution (see Grønneberg & Foldnes, 2019, for information about simulating categorical items based on non-normal distribution), and measurement model type/factor loadings. Paxton et al. (2001) underscore sample sizes below 100 as especially important for MCSS in general since they are routinely used in research. This is reflected in the fact that sample sizes below 100 are well researched in the field of reliability (e.g., Romano et al., 2010; Zimmerman et al., 1993; Zinbarg et al., 2007). Among the most prevalent factors, it appears simulee sample size, number of items, and measurement model type/factor loadings display an upward trend with medium, low to medium, and low to medium ES. The number of scale points and distribution shape display no trend (Figure 1).

Another specific characteristic of MCSS in the field of reliability is the influence of test length or number of items. If the measurement model is parallel, as assumed in most of the reviewed studies, then reliability nonlinearly increases as the number of items increases up to a certain point (Lord & Novick, 1967). Thus, the number of items is a potential confounder in the relationship between many commonly investigated factors and reliability estimates and it should be included in the design under many circumstances, as done in some studies (e.g., Komaroff, 1997; Maydeu-Olivares et al., 2007).

Furthermore, there are some specifics regarding the external validity of MCSS in the field of reliability. These specifics stem from the fact that generic reliability formula exists (Lord & Novick, 1967), the mathematical relatedness of reliability coefficients, as well as from different performance of coefficients based on the empirical conditions. As shown in Table 2, using conditions with a unidimensional parallel measurement model, uncorrelated errors, continuous scale, and normal distribution reduce the external validity overall, but the amount of reduction depends on the selected estimators. For instance, assuming a parallel measurement model may limit the external validity and conclusions for coefficients such as $\alpha$, but it can be expected that bias of coefficients such as $\omega$ would not be affected by the departure from the parallel measurement model if other varied conditions are held constant and only measurement model type is hypothetically changed. Thus, the external validity of estimators with fewer restrictions could be interpreted as somewhat greater compared to some other coefficients under identical investigated conditions, at least in terms of bias.

\textsuperscript{10} Positive bias represents overestimation, while negative bias represents underestimation of the benchmark value.
Another point to be taken into account is that if a coefficient that is, in theory, a restrictive special case of some other coefficient (e.g., K-R 20 is a special case of \( \alpha \), and \( \alpha \) is a special case of \( \omega \)) in the specified conditions is included in the study, more than one coefficient have been simultaneously investigated indirectly. Though, certain findings indicate theoretical equivalence is sometimes questionable in practical situations in the case of \( \alpha \) and \( \omega \) (e.g., Trizano-Hermosilla & Alvarado, 2016), and such findings can prompt further theoretical investigations or replication studies.

Finally, in some situations experimental control may be required. Since estimators perform differently based on measurement model type (for instance, \( \alpha \) is expectedly negatively biased if tau-equivalence is violated, but \( \omega \) is not), the researcher should decide whether to consider such factors as control and utilize parallel or tau-equivalent measurement model, depending on the research question, other factors in the design, and expected interactions. For instance, \( \alpha \) is highly dependent on the number of items (Raykov, 1997), especially if the number of items is small. However, not all the coefficients are equally influenced by test length. If the aim is to compare \( \alpha \) to another estimator which is not as dependent on the number of items as \( \alpha \) according to the previous findings, number of items could be a confounding variable irrespective of the combination of other factors in the design and it can be used as a control factor.

3.2.3. Benchmark

Benchmark was specified in most of the reviewed studies. However, the minority of the studies that used a benchmark specified more than one level for each condition (??), which suggests researchers in the field of reliability did not sufficiently capitalize on this specific MCSS advantage. On the other hand, the analysis of trends over time suggests that the use of benchmarks displays an increasing trend with medium ES (Figure 1).

Benchmark specification can be intricate in the context of reliability. There are numerous examples of different benchmark specifications. It can be specified for each condition (e.g., Shu & Schwarz, 2014), based on a certain factor in experimental design, such as distribution (e.g., Sheng & Sheng, 2012), the number of items (e.g., Trizano-Hermosilla & Alvarado, 2016), \( \alpha \) unadjusted for clustering in the case of investigating multilevel reliability (Bonito et al., 2012), or for every coefficient under a specific condition (e.g., Trizano-Hermosilla et al., 2021). Population reliability can also be one of the factors with a certain number of levels (e.g., Zumbo et al., 2007) which is desirable since population reliability, similarly as the number of items, may moderate the relationship between selected factors in experimental design and observed estimator values.

If applied to a theoretical infinite population, where the construct is unidimensional, measurement model type is parallel, and errors are uncorrelated, every internal consistency coefficient can be considered equal to population reliability. In such cases, it is justifiable to set the benchmark as a single value of population reliability (for each level if more than one is used). Due to the relationship between test length and reliability, test length is desirable as one of the factors in the study design with a unidimensional parallel measurement model. On the other hand, if a latent structure that departs from a unidimensional parallel measurement model is specified as the “correct” model or benchmark, it is possibly justifiable to set a benchmark for each coefficient as coefficient population value under each model.

The benchmark can be completely theoretical or based on an estimate in a finite simulated population. In the case of the latter, it is possible to somewhat increase the external validity for empirical populations with equal characteristics as the finite simulated population. Also, if the benchmark is based on an empirical dataset, the potential advantage is that the results
expand on the conditions encountered in practice and likely have higher external validity for the certain population and the specific instrument. However, one potential pitfall, in this case, is that if the coefficient calculated on the empirical dataset is used as the benchmark (e.g., Cuesta Izquierdo & Fonseca-Pedrero, 2014), then the benchmark itself is potentially biased if the coefficient was selected without regard for the specific conditions. The analysis of trends over time suggests that the use of benchmarks based on finite simulated population and empirical data does not display any trend over time (Figure 1).

However, since the field does not have definite guidelines in terms of satisfactory reliability, it is not a prerequisite to justify the benchmark values. Still, two approaches to the specification of population reliability to increase the external validity can be useful:

1. specification of single level of population reliability common in the field to which the researcher aims to generalize the findings and
2. specification of more than one level of population reliability with carefully selected and justified levels.

If the researcher decides to follow (i), then basing the decision on a specific empirical dataset, as done in some cases (e.g., Cuesta Izquierdo & Fonseca-Pedrero, 2014; Ten Berge & Sočan, 2004; Wagner et al., 1990), or consulting the literature in the field as Paxton et al. (2001) suggest seems reasonable. Consulting reliability generalization studies (Vacha-Haase, 1998) could also help obtain an approximation of population reliability. If the researcher decides to follow (ii), it seems reasonable to specify levels based on recommended thresholds of satisfactory reliability as observed (e.g., Kelley & Pornprasertmanit, 2016) or possibly based on quantiles of reliability estimates in reliability generalization studies.

3.24. Performance

The selection of performance measures in MCSS depends on the aims. Typically, bias is of primary interest. When comparing estimators, precision is also informative (Morris et al., 2019). If the study investigates CIs, coverage appears to be appropriate (e.g., Kelley & Pornprasertmanit, 2016). Therefore, the justification for the selection of performance measures is less important than the selection of multiple, but appropriate, performance measures. Reliability coefficients are parameter estimates and it is reasonable to assess their bias. On the other hand, precision is occasionally overlooked in reviewed studies (Table 2), and that practice can lead to misleading conclusions. If bias alone is used as a performance measure, the conclusion may be that some coefficient is the least biased under some conditions. However, if the coefficient is the least biased but relatively imprecise, then its use cannot be recommended. Thus, both bias and precision should be assessed as done in some studies (e.g., Fife et al., 2012; Hunt & Bentler, 2015), sometimes together with some other performance measures like RMSE (e.g., Thompson et al., 2010), or statistical analysis (e.g., Y. Liu et al., 2010). While descriptive and inferential statistics appear to be equally often used over time, the use of performance measures in general, appears to display an upward trend with medium ES (Figure 1).

It is desirable to report the formula even if the performance measure used in the MCSS is well-known. In some cases, different terms were used for identical performance measures. For instance, in the case of bias—average bias, empirical bias, match between theoretical and empirical value, or mean bias error were observed. Also, different terms for precision were observed—efficiency, stability, SD/variance of replicated values, and SD of bias. The term efficiency was also used for RMSE (e.g., Thompson et al., 2010).

It was suggested to specify the performance criteria of the estimator (e.g., Boomsma, 2013; Skrondal, 2000). However, there is no definite agreement on the satisfactory threshold
of reliability estimates. Thus, according to current knowledge, no absolute cut-offs for estimator performance are required or possibly justifiable (cf. Harwell, 2019) in MCSS in the field of reliability. Criteria could be specified subjectively or based on particular recommendations, but a relative comparison of estimators is likely sufficient. In some cases, conclusions based on absolute cut-offs may be misleading.

Performance evaluation can also be done using statistical analysis. On many occasions, researchers decided to evaluate performance using statistical techniques, most often ANOVA (Table 4). In some cases, descriptive statistics were used to evaluate performance (e.g., F. Gu et al., 2013; Kim et al., 2020) which can be useful as sole performance indicators, but their usefulness is increasingly limited as the design complexity increases. Therefore, reliance on descriptive statistics alone is occasionally not sufficient. ANOVA can provide objective information on main effects and interactions, regardless of design complexity. For instance, if the value of the reliability estimate is used as a DV, it can reveal which factors are more influential and how they interact as indicated by their effect sizes, as done in some studies (e.g., Bandalos & Enders, 1996). Still, such a finding is limited as it does not provide information relative to a certain benchmark. Like in some studies, if bias is used as a DV (e.g., Y. Liu & Zumbo, 2007; Y. Liu et al., 2010), it is more informative than the value of the reliability estimate used as a DV since it enables a wider scope of inference. Using coefficient values as DV is probably more appropriate when doing exploratory ANOVA to provide additional information for the specification of factor levels. Bias and precision as DVs are probably more useful when the aim is to evaluate coefficient performance under specified conditions. On the other hand, if the design is complex (e.g., four or more fully crossed factors and five or more estimators), tables with effect size values can be confusing since their interpretation is unintuitive.

Using statistical techniques without a benchmark can result in a misleading interpretation. For instance, researchers could interpret the findings as if higher reliability is more satisfactory, which may not be the case (Savalei & Reise, 2019a). Such practice enables obtaining expected coefficient values under each specific condition, inferring which factors influence particular coefficients in which direction, and observing how the factors interact. These insights are useful but they have somewhat limited practical applicability without a benchmark. When sampling error is included, coefficients can be positively or negatively biased due to chance, especially under conditions with small sample sizes in combination with an insufficient number of replications. Thus, statistical techniques cannot adequately substitute reasonably specified benchmarks.

Finally, performance evaluation is usually more comprehensive with data visualization. Visualization does not appear to be regularly used among the reviewed studies, but Figure 1 suggests it had become more prevalent and has remained approximately equally often used in the last two decades. Data visualization facilitates pattern recognition, but it can take up more space than tables and often does not display the exact values (Morris et al., 2019). However, considering the complexity of a particular design, visualization, as done in some examples (e.g., Geldhof et al., 2014; Greer et al., 2006; Zijlmans, van der Ark, et al., 2018), is generally preferred to large tables with descriptive statistics or raw parameter estimates, which can be provided in the appendix or online supplement (see Boomsma, 2013; Feinberg & Rubright, 2016; Harwell et al., 2017).

3.25. Limitations

There are several limitations of this study. First, publication bias is possibly an issue, so the sample might not be fully representative (Boulesteix et al., 2020). Therefore, findings are
limited to studies in the English language published in peer-reviewed journals. Second, since replication studies and code availability do not seem to be typical, some studies possibly have nontrivial issues that cannot be detected based on the reported information. Third, despite the sample of reviewed studies appearing to be representative of the population, it is not large and the analysis of trends over time is somewhat underpowered. Finally, IRT and GT frameworks are less prevalent than CTT in reviewed studies, so the conclusions are somewhat limited in that regard, but it possibly indicates these are overlooked in MCSS in reliability.

3.26. Conclusions and future research

There are some key takeaways regarding the application of MCSS method and implications for future research in the field of reliability. Compared to other fields, it appears describing the data generation procedure and using statistical analysis are somewhat more prevalent. Still, previously observed shortcomings from other fields are mostly echoed. More specifically, the practice would be improved if the researchers routinely reported the software, checked the data validity, provided the code or at least output, used a large number of replications with a justification, specified the benchmark, and used several appropriate approaches to performance evaluation, especially with external validity in mind. The analysis of trends over time suggests reporting the software, providing code or output, number of replications, benchmark specification, and at providing least partial justification for the selection of conditions are already on the increase.

Also, findings indicated some gaps for future research. First, some methods and estimators that can be useful under many empirical conditions are under-researched. In the case of internal consistency, these are relatively obscure forms of $\alpha$, $\lambda_2$, $\lambda_4$, $\mu$ series, Kristof’s coefficient, Feldt-Gilmer coefficient, maximal reliability, GLB, $\omega$ family, SEM-based coefficients, and internal consistency CIs. Neglected reliability estimation methods are parallel forms, test-retest, multilevel reliability, latent class-based reliability, reliability of an individual, and Bayesian reliability. The latter four especially require more attention. Second, factors that are less commonly included in the design are hierarchical structures, number of dimensions, correlated errors, manipulated true and/or error score distribution specifically, influence of missingness, and influence of outliers. Third, further research on factor interaction using more complex designs is required, especially with the aim to determine the values of certain factors which can compensate for levels of other factors that make reliability estimators more biased, and possibly imprecise. Fourth, more frequent follow-up on previous research and replications of existing studies could strongly contribute, possibly using a meta-simulation and comparing a larger number of estimators than previous studies (E. Cho, 2022), but not limited to internal consistency. Fifth, the issue of the required minimum number of replications for specific reliability estimators is unaddressed. Until this issue is resolved, the number of replications used in MCSS is recommendably as high as feasible. Finally, in a future review, when more studies are published and the practice is hopefully improved, an update on this study with a more thorough analysis of adherence to guidelines could be conducted (in which some additional features such as RNG or software version are included), preferably in the form of a systematic review.

References


Monte Carlo simulation studies of reliability in psychometrics


**Appendix A. Study protocol**

*Identification*

1. Search query A
   - Keywords: “Monte Carlo reliability”
   - Google Scholar: 1 740 000
   - Scopus: 3965
   - Web of Science: 1628

2. Search query B
   - Keywords: “simulation reliability”
   - Google Scholar: 3 150 000
   - Scopus: 21 947
   - Web of Science: 3997

3. Search query C
   - Keywords: “simulation study reliability”
   - Google Scholar: 2 970 000
   - Scopus: 6985
   - Web of Science: 2442

*Screening*

1. Studies screened: 17 067
   - Google Scholar: 3000
   - Scopus: 6000
   - Web of Science: 8067

2. Selection criteria (using titles and abstracts)
Monte Carlo simulation studies of reliability in psychometrics

- Report type: studies published in scientific journals
- Years: up to 2021 inclusive
- Language: English
- Outcomes: Monte Carlo simulation study investigating reliability estimators (CTT, GT, IRT), with or without presenting new estimators

Eligibility
1. Studies screened for eligibility based on full text: 96
2. Exclusion criteria
   - Ineligible type of reliability
   - Is not about psychometric reliability or does not investigate reliability estimators
   - Theoretical primarily
   - Not in English
3. 40 studies excluded for the following reasons:
   - Does not pertain to psychometric reliability: 3
   - Text not in English (except abstract): 5
   - Investigates rater agreement: 2
   - Investigates reliability of profiles: 1
   - Theoretical article with simulation used only to demonstrate a point (if used): 9
   - Does not investigate reliability estimators: 20

Included
1. 56 studies included
2. Additional search using reference lists of included studies
   - Additional studies included based on screening the full text: 29
   - Final sample size: 85

Coding
1. Estimator(s)
   - Investigated estimator(s)
   - Theoretical framework of investigated estimator(s)
   - If multiple simulations were conducted employing different estimators, separate coding with "/"
2. Data generation
   - Data generation procedure description (the procedure is described somewhere in the Methods section or no description is provided)
   - Model used for the data generation (FA, IRT, GT, a certain time series model, or RNG)
   - Software used for the data generation (used for the data generation specifically, not analysis or data validity check)
   - Data validity check (includes randomness check, cross-validation with another software, or model fit/convergence in the case of FA or IRT model)
   - Full code availability (code provided to enable an exact replication in the Appendix or supplement, or at least an output of the simulation run is provided)
   - If multiple simulations were conducted with differences in the data generation, separate coding with "/"
3. Replications
   - Reported number of replications
• Whether multiple numbers of replications were used (separate coding with “/” if in multiple studies)
• Justification for the number(s) of replications, code as positive if any provided

4. Conditions
• Factors in the experimental design and corresponding factor levels
• Was the selection of conditions justified (includes the justification of selection of factors to cross and the justification of factor levels, code as positive if any provided)
• If multiple simulations with different conditions were conducted, separate coding with “/”

5. Benchmark
• Benchmark, if specified
• Whether the benchmark was based on empirical data or finite simulated population
• Benchmark was coded as a factor under “Conditions” if more than one levels were specified
• If multiple simulations with different benchmarks were conducted, separate coding with “/”

6. Performance evaluation
• Descriptive statistics
• Inferential statistical techniques
• Performance measures
• Use of visualization
• If multiple simulations using different performance evaluation methods were conducted, separate coding with “/”

Appendix B. Reviewed studies

Zimmerman and Williams (1966) **Estimator(s):** Variant of Spearman-Brown formula for dependent true and error scores • **Data generation:** Data generation procedure described, –, no software reported, no data validity check, no code provided • **Replications:** 1, number not justified • **Conditions:** Number of items (10, 100), number of response choices (2, 5), selection of conditions not justified • **Benchmark:** Theoretical value obtained by Spearman-Brown formula • **Performance evaluation:** Value yielded by computer compared to benchmark, visualization not used • **Conclusion(s):** Increase in reliability with increase in number of items is observed if true and error scores are not independent.

Zimmerman (1969) **Estimator(s):** General formula for reliability of composite tests, K-R 20 and K-R 21 (Kuder & Richardson, 1937) • **Data generation:** Data generation procedure described, RNG, no software reported, no data validity check, no code provided • **Replications:** 90, number not justified • **Conditions:** 10 simulees, 10 dichotomous items under conditions of four probability patterns of giving a correct response, selection of conditions not justified • **Benchmark:** General formula for reliability of composite tests • **Performance evaluation:** Descriptive statistics (M) compared to benchmark, visualization used • **Conclusion(s):** The results were consistent with the theoretical derivation of conditions under which each of the formulas represent the reliability of a test.
Bay (1973) *Estimator(s):* 2 CI estimators for normal distribution (e.g., R. W. B. Jackson & Ferguson, 1941; Kristof, 1963), CI estimator for reliability coefficients in case of non-normal distribution • *Data generation:* Data generation procedure described, RNG, no software reported, no data validity check, no code provided • *Replications:* 2000, number not justified • *Conditions:* 30 simulees, 8 items, distribution (normal, uniform, sum of two, three or six independent uniform distributions, exponential), distribution of error scores (uniform, normal, exponential), population reliability (by fixing error variance at 4 and using observed variance of 4, 1 and 0.36), selection of conditions not justified • *Benchmark:* Expected CI value under normal distribution • *Performance evaluation:* Descriptive statistics (M, SE of the M, type I errors fixed at percent level for each tail) compared to benchmark, visualization not used • *Conclusion(s):* The effect of non-normality of the error score distribution is negligible in case of large number of items; kurtosis of the true score distribution influences the sampling distribution and SE of reliability estimates; Bay’s CI estimator is superior to other estimators.

Lissitz and Green (1975) *Estimator(s):* α, coefficient of stability, squared correlation of true and observed results • *Data generation:* Data generation procedure described, RNG, no software reported, no data validity check, no code provided • *Replications:* 100, number not justified • *Conditions:* 50 simulees, 10 items, uniform distribution, number of scale points (2, 3, 5, 7, 9, 14), average inter-item covariance (0.2, 0.5, 0.8), selection of conditions not justified • *Benchmark:* – • *Performance evaluation:* Descriptive statistics (M, SD, average increase in reliability), visualization used • *Conclusion(s):* Reliability levels off after 5 scale points.

Marshall (1976) *Estimator(s):* ANOVA-based coefficient (Hoyt, 1941) • *Data generation:* Data generation procedure described—based on a previous empirical experiment (Cotton, 1971), RNG, language of the SNOBOL family, no data validity check, no code provided • *Replications:* 150, number not justified • *Conditions:* 50 simulees, 24 trials of an experiment conducted by Cotton (1971), selection of conditions based on previous empirical study • *Benchmark:* – • *Performance evaluation:* Descriptive statistics (M, variance, skewness), visualization used • *Conclusion(s):* The implications of individual differences for the Bower-Trabasso model are obviously relevant and deserving of further attention.

Jenkins and Taber (1977) *Estimator(s):* α, coefficient of stability, squared correlation of true and observed results • *Data generation:* Data generation procedure described, RNG, no software reported, no data validity check, no code provided • *Replications:* 100, number not justified • *Conditions:* 50 simulees, uniform distribution, number of scale points (2, 3, 5, 7, 9, 10, 14), number of items (2, 3, 5, 7, 9, 10, 14), average inter-item covariance (0.2, 0.5, 0.8), simulee judgement precision (0.5, 0.7, 0.85, 1), selection of conditions not justified • *Benchmark:* – • *Performance evaluation:* Descriptive statistics (M), ANOVA with η², visualization used • *Conclusion(s):* Reliability levels off after 5 scale points, increasing number of items is useful when error variance is high.

Sedere and Feldt (1977) *Estimator(s):* α, Guttman’s λ₂, Kristof’s (1974) coefficient, Feldt’s (1975) coefficient • *Data generation:* Data generation procedure described, RNG, no software reported, no data validity check, no code provided • *Replications:* 1500, number not justified • *Conditions:* Population reliability (0.7, 0.8, 0.9), simulee sample size (100, 200, 400), part-test length ratio (1:1:1, 3:4:5, 1:2:3, 1:3:8, 1:2:9, not all levels
used for every coefficient), selection of conditions not justified • Benchmark: Population reliability per each condition of population reliability level and part-test length ratio • Performance evaluation: Descriptive statistics (M, SE of the M) compared to population reliability, visualization not used • Conclusion(s): If test parts are of unequal length, Kristof’s and Feldt’s coefficients are appropriate, if test parts are of equal length, $\lambda_2$ Kristof’s and $\alpha$ are appropriate, the empirical sampling distribution of Kristof’s, Feldt’s and $\lambda_2$ coefficients correspond to theoretical $\alpha$ distribution.

Bardo and Hughey (1978) Estimator(s): $\alpha$ • Data generation: Data generation procedure described, RNG, no software reported, no data validity check, no code provided • Replications: 100, number not justified • Conditions: 100 simulees, 5 items, distribution (normal, rectangular), selection of conditions not justified • Benchmark: $\alpha$ under normal distribution • Performance evaluation: Descriptive statistics (M, SD, minimum, maximum), $t$ test for estimator value across distributions, visualization not used • Conclusion(s): No difference was found between the distribution of $\alpha$s derived from two result distribution shapes.

Cudeck (1980) Estimator(s): K-R 20, Cliff’s (1977) coefficient, Loevinger’s (1948) index of homogeneity • Data generation: Data generation procedure described, three-parameter (3-PL) IRT model using method of Kinderman and Ramage (1976), IBM subroutine BDTR RANDU (FORTRAN), no data validity check, no code provided • Replications: 10, number not justified • Conditions: 100 simulees, number of items (20, 40), mean item discrimination (0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2), item difficulty SD (1, 1.5), mean chance of success due to guessing (0, 0.2), mean ability (0, 0.5), selection of conditions not justified • Benchmark: $\alpha$ under normal distribution • Performance evaluation: ANOVA, visualization used • Conclusion(s): In case of short tests or tailored testing, Cliff’s coefficient is appropriate, otherwise K-R 20 is appropriate.

Huynh (1986) Estimator(s): K-R 20 estimated by three methods: ANOVA, fitting constants (FITCO; Henderson, 1953), symmetric sums (SYSUM; Koch, 1968) • Data generation: Data generation procedure described—generation based on the IRT model / random sampling without replacement from three empirical datasets, beta-binomial IRT model, no software reported, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: 40 simulees, 20 dichotomous items, population reliability (0.5, 0.7, 0.9), true ability distribution skewness (0, ±0.386, ±0.519, ±0.9), proportion of missing data (0.2, 0.4) / 138 items, simulee sample size (528, 428, 314), skewness (0.259, 0.096, −0.271), proportion of missing data (0.2, 0.4), selection of conditions not justified • Benchmark: Population reliability • Performance evaluation: Descriptive statistics (M), MSE, visualization not used • Conclusion(s): ANOVA method of K-R 20 estimation is the most appropriate in case of missing data.

Sijtsma and Molenaar (1987) Estimator(s): Mokken’s (1971) Method 1 and Method 2, Molenaar and Sijtsma’s (1984) and Sijtsma and Molenaar’s (1987) 2 methods, $\alpha$, $\mu_2$ and $\mu_3$ (Ten Berge & Zegers, 1978) • Data generation: Data generation procedure described, two-parameter (2-PL) IRT model, no software reported, partial data validity check—some unspecified conditions were cross-validated, no code provided • Replications: 200, number not justified • Conditions: Item difficulty for each number of items (0.2, 0.67 / 0.1, 0.3), item discrimination (1, 3), simulee sample size (100, 300), number of items (7, 15), selection of conditions partially justified—sample size and number of items levels justified • Benchmark: Population reliability, specified for
each condition. **Performance evaluation:** SD, RMSE, average bias, visualization not used. **Conclusion(s):** Mokken’s (1971) Method 1 and Molenaar and Sijtsma’s (1984) and Sijtsma and Molenaar’s (1987) two methods are the least biased, $\alpha$, $\mu_2$ and $\mu_3$ display slightly higher stability than IRT-based methods.

Wagner et al. (1990) **Estimator(s):** Odd-even split-half coefficient, $\alpha$, Guttman’s $\lambda_4$ • **Data generation:** Data generation procedure described—based on previous Rorschach test project empirical dataset, RNG, no software reported, no data validity check, no code provided • **Replications:** 100, number not justified • **Conditions:** 50 simulees, selection of conditions not justified • **Benchmark:** Theoretical distribution of $\alpha$ (Feldt et al., 1987) • **Performance evaluation:** Descriptive statistics (M, SD, median, mode, minimum, maximum, skewness, kurtosis) for comparison of estimators to each other, percentiles of $\alpha$ to compare it to theoretical distribution, visualization not used • **Conclusion(s):** Guttman’s $\lambda_4$ yields the highest value and is the most stable, $\alpha$ appears to be a slightly inferior alternative, whereas odd-even split-half coefficient is inferior to other two coefficients; empirical distribution of $\alpha$ is similar to theoretical.

Smith and Luecht (1992) **Estimator(s):** G coefficient • **Data generation:** Data generation procedure described, single facet cross-design with lag-1 dependence among facet effects as serial correlation, no software reported, no data validity check, no code provided • **Replications:** 1000, number not justified • **Conditions:** Number of levels of the facet (10, 25, 50), simulee sample size (10, 25, 50), lag-1 serial correlations (0.2, 0.4, 0.6, 0.8) / 25 facet levels, simulee sample size (2, 3, 4, 5) for lag-1 serial correlation 0.2, selection of conditions not justified • **Benchmark:** – • **Performance evaluation:** M variance component estimate (random effects ANOVA), bias introduced into variance component estimates, visualization not used • **Conclusion(s):** Bias introduced by the serial correlation is a function of the number of levels of the facet represented in the G study design and of the relative magnitude of the population values of the variance components.

Zimmerman et al. (1993) **Estimator(s):** $\alpha$ • **Data generation:** Data generation procedure described, RNG, no software provided, data validity checked by using three methods (Box & Muller, 1958; Marsaglia & Bray, 1964; central limit theorem method) to check for potential differences between methods, no code provided • **Replications:** 2000, number not justified • **Conditions:** 10 items, population reliability (0.6, 0.9), simulee sample size (20, 80) / 5 items, population reliability 0.75, simulee sample size (5, 10, 20, 40, 80, 160) / 40 simulees, 10 items, population reliability (0.6, 0.75, 0.9), distribution (normal, uniform, exponential, mixed-normal) / 10 simulees, 10 items, population reliability 0.8, true score matrix (additive, multiple non-additive, exponential non-additive and random non-additive, non-additive representing congeneric measurement model) / 20 simulees, 8 items, population reliability (0.5, 0.6, 0.7, 0.8, 0.9), true score matrix (additive, multiple non-additive, exponential non-additive and random non-additive) / 20 simulees, 10 items, population reliability 0.80 error correlatedness (0, 0.25, 0.4), number of correlated errors (3, 6) / 20 simulees, 8 items, population reliability (0.8, 0.5), error correlatedness (0.1, 0.3, 0.5), number of pairwise correlated errors (2, 3, 4, 5, 6, 7, 8), selection of conditions not justified • **Benchmark:** Population reliability, specified for each condition • **Performance evaluation:** Descriptive statistics (relative frequencies, M, SD) compared to population reliability, visualization not used • **Conclusion(s):** Reported values of $\alpha$ may be both far above and far below the population reliability.
Bacon et al. (1995) *Estimator(s):* \( \alpha \), unit-weighted \( \omega \), \( \omega \) with unequal weights • *Data generation:* Data generation procedure not described, FA model, no software reported, no data validity check, no code provided • *Replications:* Not reported • *Conditions:* Loadings range (0.3–0.9, 0.4–0.9, 0.5–0.9, 0.6–0.9, 0.7–0.9, 0.8–0.9), loadings skewness (0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2), number of items (3–15), selection of conditions partially justified—loadings range and number of item levels justified • *Benchmark:* • *Performance evaluation:* Descriptive statistics (M) / type I specification error investigated by examining points of item deletion in which reliability of composite with \( j \) items becomes equal to reliability of composite with \( j - 1 \) items, used visualization • *Conclusion(s):* Differences between selected coefficients were more prominent in case of shorter tests and higher dispersion of loadings.

Bost (1995) *Estimator(s):* G and D coefficients obtained by ordinary least squares (OLS) ANOVA in generalizability analysis • *Data generation:* Data generation procedure described, RNG, PROC IML of SAS, no data validity check, no code provided • *Replications:* 50, number justified—“to adequately account for the variability inherent in the simulation process” • *Conditions:* Simulee sample size per occasion (15, 25, 50), number of occasions (3, 5, 7), error structures (first-order stationary autoregressive, first-order nonstationary autoregressive), 6 variance component combinations, selection of conditions partially justified—simulee sample size, number of occasions, variance component combinations levels justified • *Benchmark:* True variance values for each variance component combination condition • *Performance evaluation:* Percent difference (relative bias), used visualization • *Conclusion(s):* Traditional G and D coefficients are not often not appropriate when errors are correlated.

Bandalos and Enders (1996) *Estimator(s):* \( \alpha \) • *Data generation:* Data generation procedure described, RNG, SAS, no data validity check, no code provided • *Replications:* 1000, number not justified • *Conditions:* 100 simulees, 10 items, distribution (normal, 2 skewed, platykurtic and leptokurtic), inter-item correlation (0.25, 0.5, 0.75), number of scale points (3, 5, 7, 9, 11), selection of conditions not justified • *Benchmark:* • *Performance evaluation:* Descriptive statistics (M), ANOVA with \( \eta^2 \), visualization not used • *Conclusion(s):* Researchers should consider the underlying distribution shape of measured constructs, \( \alpha \) levels off after 5 or 7 scale points.

Barchard and Hakstian (1997b) *Estimator(s):* 2 \( \alpha \) CI estimators (Feldt, 1965; Hakstian & Whalen, 1976) • *Data generation:* Data generation procedure described, RNG, no software reported, data validity partially checked—a new set was generated if a generated set of variances contained one or more that was negative, no code provided • *Replications:* 20 000 / 5000, number partially justified—“to ensure that the slightest departure from nominal levels would be detected” for the first number • *Conditions:* 100 simulees, number of items (5, 20), population reliability (0.6, 0.75, 0.9), type 12 sampling, measurement model type (compound symmetrical/parallel, spherical) / 100 simulees, number of items (5, 20), population reliability (0.6, 0.75, 0.9), type 12 sampling, measurement model type (parallel, essentially parallel, tau-equivalent, essentially tau-equivalent), selection of conditions not justified • *Benchmark:* Population reliability / \( \alpha \) • *Performance evaluation:* Proportion of CIs that included population reliability (coverage), aggregated proportion of CIs that included population reliability, frequency and proportion of CI values not spanning the parameter, visualization not used • *Conclusion(s):* No performance differences between the two estimators were
Monte Carlo simulation studies of reliability in psychometrics

found.

Barchard and Hakstian (1997a) **Estimator(s)**: 2 \( \alpha \) CI estimators (Feldt, 1965; Hakstian & Whalen, 1976) • **Data generation**: Data generation procedure described, RNG, no software reported, data validity checked—error variances and target population value of \( \alpha \) outside of specified ranges resulted in new generation and overall validity check, no code provided • **Replications**: 20,000 / 5,000, number partially justified—“because Feldt (1965) and Hakstian and Whalen (1976) had seen the differences between type 1 and type 12 sampling as small” for the first number • **Conditions**: 100 simulees, number of items (5, 20), population reliability (0.6, 0.75, 0.9), sampling type (1, 12), measurement model type (compound symmetrical/parallel, tau-equivalent, spherical), selection of conditions not justified • **Benchmark**: Population reliability / \( \alpha \) • **Performance evaluation**: Proportion of CIs that spanned population reliability (coverage), visualization not used • **Conclusion(s)**: The results indicate that if the measurement model is in line with essential parallelism, the two methods maintain type 1 error control; no difference in performance between the methods was observed.

Komaroff (1997) **Estimator(s)**: \( \alpha \), \( \alpha \)-corrected for error covariances (Komaroff, 1997) • **Data generation**: Data generation procedure described, RNG, RANNOR function in SAS, no data validity check, no code provided • **Replications**: 150, number not justified • **Conditions**: 10,000 simulees, number of items (6, 12, 18), inter-item correlation (0.2, 0.5, 0.7), error correlatedness (0, 0.2, 0.5, 0.7), number of items with correlated errors (3 for 6, 12, and 18 items, 6 for 12 and 18 items, 9 for 18 items), measurement model type (parallel, tau-equivalent), true score correlation (0, 0.2, 0.5, 0.7, 1), selection of conditions not justified • **Benchmark**: Population reliability for each number of items and true score correlation condition • **Performance evaluation**: Descriptive statistics (M) compared to population reliability, visualization not used • **Conclusion(s)**: CFA error covariance can be subtracted from \( \alpha \) to substantially reduce or eliminate the positive bias resulting from positively correlated errors.

Barnette (1999) **Estimator(s)**: \( \alpha \) • **Data generation**: Data generation procedure described, RNG, program written in QuickBASIC, data validity partially checked using PROC CORR ALPHA of SAS—population \( \alpha \) values checked, no code provided • **Replications**: 50, number not justified • **Conditions**: 100 simulees, 50 items with 7 scale points, uniform distribution, population reliability (0.7, 0.8, 0.9), percentage of randomly replaced data due to missing data (5 %, 10 %, 15 %, 20 %), population reliability (0.7, 0.8, 0.9), 8 responding patterns, selection of conditions not justified • **Benchmark**: Population reliability / \( \alpha \) • **Performance evaluation**: Descriptive statistics (M) compared to population reliability, visualization not used • **Conclusion(s)**: Findings illustrate that different patterns have differential systematic effects and that even relatively few occurrences of some patterns may highly influence \( \alpha \).

Enders and Bandalos (1999) **Estimator(s)**: \( \alpha \) • **Data generation**: Data generation procedure described, RNG, RANNOR function within the IML of SAS, no data validity check, no code provided • **Replications**: 500, number not justified • **Conditions**: 100 simulees, 12 items, average inter-item correlation (0.25, 0.5, 0.75), baseline distribution (normal, moderately skewed, highly skewed and leptokurtic), contrast distribution (normal, moderately skewed, highly skewed and leptokurtic), number of items with contrasting distribution (1, 3, 6), number of scale points (3, 5, 7, 9), selection of conditions not justified • **Benchmark**: – • **Performance evaluation**: Descriptive statistics (M, SD),
ANOVA with partial $\eta^2$, visualization not used • Conclusion(s): Decrease in $\alpha$ is a complex function of interactions among distribution shape, number of scale points and average inter-item correlation; adding more scale categories compensates the loss of variance due to non-normality; various distribution shapes only slightly and similarly affect internal consistency.

Sideridis (1999) Estimator(s): $\alpha$ • Data generation: Data generation procedure described, RNG, input programs written in SPSS/PC+, no data validity check, no code provided • Replications: 100 / 200 for U-shaped distribution, number not justified • Conditions: 250 simulees, 10 items, distribution (normal, uniform, skewed, leptokurtic, U-shaped), selection of conditions partially justified—number of simulees and number of items justified • Benchmark: – • Performance evaluation: Descriptive statistics (M, SD), ANOVA with Tukey's post hoc test, visualization not used • Conclusion(s): $\alpha$ can be used with data having various distributional properties.

Shevlin et al. (2000) Estimator(s): $\alpha$ • Data generation: Data generation procedure not described, FA model, SPSS, no data validity check, no code provided • Replications: 50, number not justified • Conditions: 6 items, factor loadings (0.3, 0.5, 0.7), error correlatedness between 2 items (0, 0.1, 0.2, 0.3), simulee sample size (50, 100, 200, 400), selection of conditions not justified • Benchmark: Population reliability based on population item-covariance matrix in accordance with specified factor loadings and error correlatedness and 50 000 simulees • Performance evaluation: Descriptive statistics (M, SD), Levene’s test of homogeneity of variances, ANOVA, visualization used • Conclusion(s): Correlated errors increase the value of $\alpha$.

Osburn (2000) Estimator(s): $\alpha$, standardized $\alpha$, stratified $\alpha$ (Cronbach et al., 1965), Feldt’s (1975) coefficient, Gilmer and Feldt’s (1983) coefficient, Guttman’s $\lambda_2$ and $\lambda_4$, Kristof’s (1974) coefficient, Raju’s (1982) coefficient, maximal reliability with component weighting procedure • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: Not reported • Conditions: 4 items, number of dimensions (1, 2), measurement model type (parallel, tau-equivalent, congeneric, correlation between dimensions (0.8, 0.4, 0.2) / 8 items, number of dimensions (1, 2, 3), measurement model type (parallel, tau-equivalent, congeneric, correlation between dimensions (0.8, 0.4, 0.2), selection of conditions not justified • Benchmark: Population reliability specified for each measurement model type and correlation between dimensions condition • Performance evaluation: Descriptive statistics (M) compared to population reliability, visualization not used • Conclusion(s): Guttman’s $\lambda_4$ is the most accurate estimator of population reliability in all conditions.

Dimitrov (2003) Estimator(s): IRT-based marginal reliability • Data generation: Data generation procedure described, 2-PL IRT model, SAS, data validity checked with XCAL-IBRE, no code provided • Replications: Not reported • Conditions: Congeneric measurement model, 8000 simulees, 20 dichotomous items, selection of conditions not justified • Benchmark: Theoretical IRT-based coefficient value • Performance evaluation: Match between theoretical and empirical value, visualization not used • Conclusion(s): The proposed formulas have methodological and computational value in bridging concepts of IRT and CTT.
Duhachek and Iacobucci (2004) *Estimator(s):* $\alpha$ CI estimator (Duhachek & Iacobucci, 2004) / $\alpha$ CI estimator (Duhachek & Iacobucci, 2004) / six $\alpha$ CI estimators (e.g., Duhachek & Iacobucci, 2004; Feldt, 1965; Hakstian & Whalen, 1976) • *Data generation:* Data generation procedure described (except for the first study), RNG, no software reported, no data validity check, no code provided • *Replications:* Not reported / 1000 / 1000, numbers not justified • *Conditions:* Simulee sample size (30, 50, 100, 200), number of items (2, 3, 5, 7, 10), average inter-item correlation (0, 0.1, 0.2, ..., 1), item SD (1–2, 1–5) / simulee sample size (30, 50, 100, 200), number of items (4–10), average inter-item correlation (0.3, 0.4, ..., 0.7), covariance heterogeneity (parallel measurement model, two heterogeneous inter-item correlation matrices) / simulee sample size (30, 50, 100, 200), number of items (5, 7), average inter-item correlation (0.4, 0.5, 0.6, 0.7), covariance heterogeneity (the unidimensional case representing compound symmetry and a multidimensional case representing two underlying factors), selection of conditions partially justified (only in the first study)—simulee sample size and number of items levels justified • *Benchmark:* Population reliability / $\alpha$ • *Performance evaluation:* SE, multiple regression with SE of $\alpha$ as DV / SE, ANOVA with $\alpha$ and SE as DVs / coverage, CI width, visualization used • *Conclusion(s):* The $\alpha$ CI estimator (Duhachek & Iacobucci, 2004) is superior to other estimators.

Ten Berge and Šočan (2004) *Estimator(s):* $\alpha$, Guttman’s $\lambda_4$, GLB, FA-based reliability coefficient (Ten Berge & Šočan, 2004) • *Data generation:* Data generation procedure described—based on an empirical dataset (De Leeuw, 1983), RNG, software not reported, no data validity check, no code provided • *Replications:* 500, number not justified • *Conditions:* 6 items, simulee sample size (100, 250, 500, 1,000), selection of conditions not justified • *Benchmark:* Population reliability, based on value of each coefficient in the empirical data of a questionnaire • *Performance evaluation:* Empirical bias, visualization not used • *Conclusion(s):* The sampling bias problem of the GLB does not play a role when the number of test parts is small, as is often the case with congeneric measures. GLB and unidimensionality-based reliability are often equal when there are three test parts.

Christmann and Van Aelst (2006) *Estimator(s):* $\alpha$, 4 robust variants of $\alpha$ (Christmann & Van Aelst, 2006; Wilcox, 1992) • *Data generation:* Data generation procedure described, RNG, no software reported, no data validity check, no code provided • *Replications:* 1000, number not justified • *Conditions:* Simulee sample size (40, 100, 500), number of items (2, 10), inter-item correlation (0, 0.2, 0.5, 0.8), proportion of contamination with outliers (0.05, 0.1, 0.2), probability model (multivariate normal, multivariate Student’s $t$ with $df = 3$, contamination model with one different covariance matrix, contamination model with different location parameter and covariance matrix, contamination model with different location parameter), selection of conditions not justified • *Benchmark:* Population reliability / $\alpha$ specified for each condition • *Performance evaluation:* Bias, RMSE, visualization used • *Conclusion(s):* Robust $\alpha$ based on all three robust covariance estimators yield more stable estimates than the classical approach.

Greer et al. (2006) *Estimator(s):* Standardized $\alpha$ • *Data generation:* Data generation procedure described, RNG, subroutine RNNOF from the International Mathematical and Statistical Library for use with FORTRAN compiler, data validity checked with SAS, no code provided • *Replications:* 10,000, number not justified • *Conditions:* 100 simulees,
continuous scale, population inter-item correlation (0.1, 0.3, 0.5, 0.7), number of items (6, 10, 20), extent of item skew (±2.7, ±4.4, ±5.9), direction of item skew when items are divided into two equally large groups (same, opposite) / 100 simulees, Likert items with 5 scale points, population inter-item correlation (0.1, 0.3, 0.5, 0.7), number of items (6, 10, 20), extent of item skew (±0.6, ±1.3, ±2.2, ±3.5), direction of item skew when items are divided into two equally large groups (same, opposite), selection of conditions not justified • Benchmark: – • Performance evaluation: Descriptive statistics (M), visualization used • Conclusion(s): Results indicated that skewness decreased the average inter-item correlation and produced small decreases in $\alpha$ that were the largest when skewness was extreme, inter-item correlation was small, items were skewed in opposite directions, and there were fewer items.

Zinbarg et al. (2006) Estimator(s): Methods of estimating $\omega$ based on principal factors (PF) and principal components (PC) with higher order structure (HO) or hierarchical structure in which items loads on at least one group factor beside general: 2 methods derived from EFA (first PF and HO-PF), 2 methods derived from CFA (HO-CF and Hi-CF), 2 methods derived from PCA (first PC and HO-PC), $\alpha$ • Data generation: Data generation procedure described, FA model, EQS, data validity partially checked—model fit was examined and cross-validation was done based on 10 replications of an empirical dataset Anxiety Sensitivity Index (Reiss et al., 1986) for some conditions, no code provided • Replications: 160 / 50 additional replications for conditions with higher order factor absent / 40, numbers not justified • Conditions: 4 types of higher order factor structure (higher order factor absent, weak higher order factor loadings, mixed higher order factor loadings, strong higher order factor loadings), number of items divided into four equal-sized sets (12, 20), simulee sample size (100, 200), selection of conditions not justified • Benchmark: Population reliability / $\omega_{h}$ • Performance evaluation: Mean difference between estimator and benchmark, efficiency (SD of estimator values), correlation of estimators and benchmark, MANOVA with planned contrasts and $\eta^{2}$, visualization not used • Conclusion(s): It is recommended that the HO-CF and/or the Hi-CF methods be used to estimate $\omega_{h}$ whenever the investigator has a clear a priori measurement model that can be tested by CFA, otherwise HO-PF or HO-PC are recommended.

Walker (2006) Estimator(s): Spearman-Brown formula, Flanagan-Rulon formula • Data generation: Data generation procedure not described, RNG, no software reported, no data validity check, no code provided • Replications: Not reported • Conditions: Ratio of variances of halves (1, 1.1, 1.2, ..., 2), correlatedness of halves (1, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05), selection of conditions partially justified—elaborates upon the work of Cronbach (1951) and Charter (1996) • Benchmark: – • Performance evaluation: Spearman-Brown to Flanagan-Rulon formula values and their ratio, visualization used • Conclusion(s): Use of the Spearman-Brown formula is not warranted when the ratio between the SD on two halves of a test is disparate, or beyond 0.9 to 1.1 and in these cases Flanagan-Rulon formula should be used.

Y. Liu and Zumbo (2007) Estimator(s): $\alpha$ • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: 100, number not justified • Conditions: 14 items, simulee sample size (50, 100, 200, 500, 1000), population reliability (0.4, 0.6, 0.8, 0.9), outlier contamination proportion (1 %, 8 %, 15 %), mean shift of the contamination distribution (0, 1.5, 3), SD of
the contamination distribution (1, 1.5, 3), selection of conditions partially justified—levels of the independent variables were selected to represent those found in the literature • Benchmark: Population reliability • Performance evaluation: Bias, efficiency (MSE), ANOVA with bias and efficiency as DVs and $\eta^2$; visualization used • Conclusion(s): $\alpha$ is not affected by symmetric outlier contamination, whereas asymmetric outliers artificially inflate the estimates of coefficient $\alpha$; $\alpha$ estimates are positively biased and more variable with increasing asymmetry and proportion of outlier contamination.

Maydeu-Olivares et al. (2007) Estimator(s): Asymptotically distribution-free (ADF; Maydeu-Olivares et al., 2007) and normal-theory (NT) $\alpha$’s CI estimators • Data generation: Data generation procedure described, RNG, no software reported, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: Parallel measurement model, simulee sample size (50, 100, 200, 400), number of items (5, 20), factor loadings (0.4, 0.6, 0.8), item type in terms of number of scale points and skewness (three with 2 scale points, three with 5 scale points, varying in skewness and/or kurtosis), population reliability (0.25–0.77, 0.49–0.91, 0.73–0.97) / congeneric measurement model, simulee sample size (50, 100, 400, 1000), number of items (7, 14, 21), item type in terms of number of scale points and skewness (three with 2 scale points, three with 5 scale points, varying in skewness and/or kurtosis), selection of conditions partially justified—simulee sample size, number of items, and factor loadings levels justified • Benchmark: Population reliability / $\alpha$ / population reliability / $\omega$ • Performance evaluation: Relative bias of SE, coverage, visualization used • Conclusion(s): NT CIs can be used when items are approximately normally distributed even if sample size is small; for sample sizes over 100, the ADF CIs provide an accurate estimate of population $\alpha$, for sample sizes over 100, they are preferred to NT CIs.

Zinbarg et al. (2007) Estimator(s): 6 estimators of $\omega_h$ using principal factors (PF) or confirmatory (CF) method with equal higher-order loadings (EQ) or maximally different higher-order loadings (MD): (first PF, higher-order PF EQ, higher-order PF MD, higher-order CFEQ, higher-order CFMD), $\alpha$ • Data generation: Data generation procedure described, RNG, EQS, data validity partially checked using confirmatory fit index (CFI), no code provided • Replications: 10, number not justified • Conditions: 10 items, two dimensions, ratio of covariance between any two items loading on different group factors and correlation between any two items loading on the same group factor (0.1, 0.3, 0.5 / 0.05, 0.15, 0.25), simulee sample size (25, 50, 100, 200, 400, 800), selection of conditions partially justified—two dimensions and simulee sample size levels selected to examine a broader range of conditions compared to Zinbarg et al. (2006) • Benchmark: Population reliability / $\omega_h$ • Performance evaluation: Descriptive statistics (M, efficiency (SD of estimator values)), correlation of estimators and benchmark, repeated measures MANOVA with bias as DV and planned contrasts, visualization not used • Conclusion(s): Range of possible $\omega_h$ produced by either the higher-order PF or higher-order CF methods will be less positively biased than $\alpha$ or the first PF method of estimating $\omega_h$ in estimating the internal consistency of a scale containing two group factors with an underlying general factor.

Zumbo et al. (2007) Estimator(s): $\alpha$, ordinal $\alpha$ based on polychoric correlation, ordinal theta coefficient (Zumbo et al., 2007) • Data generation: Data generation procedure described, FA model, SPSS, no data validity check, no code provided • Replications: 10 000, number
not justified • **Conditions**: 14 items, population reliability / $\alpha$ (0.2, 0.4, 0.6, 0.8, 0.9), item skewness (0, $-2$) number of scale points (2–7), selection of conditions partially justified—number of items justified • **Benchmark**: Population reliability • **Performance evaluation**: Descriptive statistics (M) compared to population reliability, visualization not used • **Conclusion(s)**: Ordinal $\alpha$ and theta are appropriate alternatives to $\alpha$ with Likert response data.

Laenen et al. (2008) **Estimator(s)**: Reliability coefficient ($R_A$) for an entire longitudinal sequence (Laenen et al., 2008), reliability coefficient ($R_T$) as an average reliability across different time points (Laenen et al., 2007) • **Data generation**: Data generation procedure described, linear mixed model with random intercept without and with random slope, no software reported, no data validity check, no code provided • **Replications**: 500, number not justified • **Conditions**: Model type (random intercept, random slope), simulee sample size (50, 150), proportion of error variance (0.09, 0.5, 0.9), number of time points (3, 6, 9), selection of conditions not justified • **Benchmark**: – • **Performance evaluation**: Descriptive statistics (M, CI calculated using matrix differential calculus combined with the delta method), visualization used • **Conclusion(s)**: $R_T$ and $R_A$ should be interpreted in different ways.

Lozano et al. (2008) **Estimator(s)**: $\alpha$ • **Data generation**: Data generation procedure described, FA model, PRELIS, no data validity check, no code provided • **Replications**: 100, number not justified • **Conditions**: 30 items, average inter-item correlation (0.2, 0.3, ..., 0.9), simulee sample size (50, 100, 200, 500), number of scale points (2, 3, ..., 9), selection of conditions not justified • **Benchmark**: – • **Performance evaluation**: Descriptive statistics (M, coefficient of variation), visualization used • **Conclusion(s)**: The optimum number of alternatives is between four and seven.

Yurdugül (2008) **Estimator(s)**: $\alpha$ • **Data generation**: Data generation procedure described, RNG, SIMREL software written in Visual Basic, no data validity check, no code provided • **Replications**: 100, number not justified • **Conditions**: Simulee sample size (30, 100, 300, 500), number of items (5, 6, ..., 20), first eigenvalue size (small, moderate, large), selection of conditions not justified • **Benchmark**: $\alpha$ in 10 000 populations, each consisting of 5000 simulees • **Performance evaluation**: Relative bias, relative RMSE, ANOVA with relative bias and relative RMSE as DVs and $\omega$ effect size, visualization used • **Conclusion(s)**: If the the first eigenvalue is higher than 6, the sample $\alpha$ is an especially robust estimator of the population $\alpha$ even when $n = 30$; if the first eigenvalue is between 3 and 6, $n = 100$ is sufficient for an unbiased estimate of $\alpha$.

Laenen et al. (2009) **Estimator(s)**: SEM-based coefficient of stability • **Data generation**: Data generation procedure described, linear mixed model (LMM) with random intercept without and with random slope, no software reported, no data validity check, no code provided • **Replications**: 250, number not justified • **Conditions**: 5 time points, error variability (30, 300, 3000), simulee sample size (50, 150), proportion of error variance (9 %, 50 %, 90 %), selection of conditions not justified • **Benchmark**: – • **Performance evaluation**: Descriptive statistics (M, CI calculated like in Laenen et al., 2007), coverage based on restricted maximum likelihood, visualization not used • **Conclusion(s)**: The method is recommended for use in practice in case of measurement at multiple time points.
Kim and Feldt (2010) estimator(s): IRT-based coefficient / IRT-based coefficient, $\alpha$, Gilmer and Feldt’s (1983) coefficient • Data generation: Data generation procedure described, 3-PL IRT model / 3-PL IRT and generalized partial credit model with random samples from an empirical dataset of large-scale test battery for middle school students, PARSCALE (for the second study), no data validity check, no code provided • Replications: Not reported / 30, number not justified • Conditions: Parallel measurement model, 18 dichotomously scored multiple-choice item types with combinations of crossing three IRT $a$ (0.5, 1, 1.5), three $b$ (−1, 0, 1) and two $c$ (0, 0.2) parameters, number of items (10, 30, 50) / congeneric measurement model, 9 subtests representing different item heterogeneity, simulee sample size (200, 1000, 5000), selection of conditions not justified • Benchmark: Population reliability / $\alpha$ • Performance evaluation: Descriptive statistics (M) / descriptive statistics (M, SD), visualization not used • Conclusion(s): IRT-based coefficient outperforms CTT based coefficients and is closer to Gilmer and Feldt’s (1983) coefficient than $\alpha$; to increase reliability, items with high discrimination and difficulty that matches sample average ability should be used.

Y. Liu et al. (2010) estimator(s): $\alpha$ • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: 100, number not justified • Conditions: 14 items, 100 simulees, contamination proportions (1 %, 8 %, 15 %), mean shift of the sampling from the contamination distribution (0, 1.5, 3), population reliability (0.4, 0.6, 0.8, 0.9), number of scale points (2, 3, ..., 7), selection of conditions not justified • Benchmark: Population reliability / $\alpha$ • Performance evaluation: ANOVA with bias and efficiency (MSE) as DVs and $\eta^2$, visualization used • Conclusion(s): Outliers can inflate the estimates of $\alpha$ in case of ordinal response scales.

Romano et al. (2010) estimator(s): 8 CI estimators of $\alpha$ (e.g., Bonett, 2002; Fisher, 1950; Hakstian & Whalen, 1976; Maydeu-Olivares et al., 2007) • Data generation: Data generation procedure described, 3-PL IRT model, RANNOR function in SAS/IML, data validity partially checked—by hand-checking results from benchmark data sets, no code provided • Replications: 10,000, number justified—provides adequate precision for coverage CI and width (Robey & Barcikowski, 1992) • Conditions: Dichotomous items, population reliability (0.5, 0.7, 0.9), sample size (10, 50, 100, 1000), number of items (5, 10, 20, 40), selection of conditions not justified • Benchmark: Population reliability / $\alpha$ based on population of 50,000 simulees • Performance evaluation: CI coverage, CI width, ANOVA with bias in point estimates, coverage rate and width as DVs, visualization used • Conclusion(s): The estimators that proved to be the most accurate were those proposed by Bonett (2002) and Fisher (1950) but the differences among all estimators are negligible in most conditions.

Thompson et al. (2010) estimator(s): $\alpha$, Kristof’s (1963) coefficient, Guttman’s maximal $\lambda_4$, $\lambda_4$-corrected, $\lambda_4$, cross-validated $\lambda_4$, corrected cross-validated • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: 2000 / 250 for conditions with 16 items, second number justified—due to time length required to calculate maximal $\lambda_4$ • Conditions: Simulee sample size (20, 40, 60, 100, 200), number of items (4, 6, 8, 10, 16), measurement model type (3 levels of tau-equivalent model, 2 levels for congeneric model, 3 levels for multifactor hierarchical structure), selection of conditions par-
Novak and Rebernjak

tially justified—measurement model type levels justified • Benchmark: Population reliability specified for each measurement model type and number of items condition • Performance evaluation: Bias, efficiency (SD of estimators), precision (RMSE), visualization used • Conclusion(s): $\alpha$ is a poor estimator of internal consistency unless tau-equivalence assumption is fulfilled, maximal $\lambda_4$ mostly produces positively biased estimates, corrected cross-validated maximal split-half coefficient was in most conditions relatively unbiased but displayed poor efficiency.

Romano et al. (2011) Estimator(s): 8 CI estimators of $\alpha$ (e.g., Bonett, 2002; Fisher, 1950; Haks-tian & Whalen, 1976; Maydeu-Olivares et al., 2007) • Data generation: Data generation procedure described, RNG, SAS/IML RANUNI function, no data validity check, no code provided • Replications: 10 000, number justified—provides adequate precision for coverage CI and width (Robey & Barcikowski, 1992) • Conditions: Population reliability (0.5, 0.7, 0.9), simulee sample size (10, 50, 100, 1000), number of items (5, 10, 20, 40), number of scale points (5, 7), population distribution shape (uniform, unimodal symmetrical, highly skewed), heterogeneity of inter-item covariance (0, 0.1, 0.3), selection of conditions partially justified—levels of each factor were selected to span a relatively broad range of conditions that may be encountered in applied research • Benchmark: Population reliability / $\alpha$ based on 1 000 000 simulees • Performance evaluation: Bias, coverage, average confidence band width, ANOVA with $\eta^2$, visualization used • Conclusion(s): The differences in CI estimator performance are negligible in most of the conditions; CI estimator (Maydeu-Olivares et al., 2007) is the least accurate.

Van Der Ark et al. (2011) Estimator(s): $\alpha$, Guttman’s $\lambda_2$, IRT-based reliability coefficient (MS; Molenaar & Sijtsma, 1988), random splits split-half coefficient with Spearman-Brown formula, latent class reliability coefficient (LCRC; Van Der Ark et al., 2011) • Data generation: Data generation procedure described, IRT multidimensional graded response model, R, no data validity check, code available upon request • Replications: 1000, number not justified • Conditions: Number of items (6, 18), simulee sample size (200, 1000), item format (dichotomous, polytomous with 5 categories), discrimination parameter (different or equal across items), number of dimensions (1, 2), selection of conditions not justified • Benchmark: Population reliability based on 1 000 000 simulees • Performance evaluation: Absolute bias, mean absolute error, visualization not used • Conclusion(s): LCRC is superior to other coefficients, $\lambda_2$ and MS are viable alternatives if items are unidimensional with equal discrimination, $\alpha$ and random splits-split half coefficient are not recommended due to high bias and high error.

Bonito et al. (2012) Estimator(s): Multilevel reliability coefficients (individual- and group-level) • Data generation: Data generation procedure described—based on empirical dataset (H. S. Park, 2008), RNG, no software reported, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: Number of items (5, 10), number of clusters (2, 4, 6), selection of conditions not justified • Benchmark: Raw $\alpha$ estimate unadjusted for clustering • Performance evaluation: Descriptive statistics (M) compared to benchmark, visualization not used • Conclusion(s): For six-person clusters, a moderate to small correlation among items provides acceptable group-level reliability estimates in both 5- and 10-item scales; for two-person groups, a large inter-item correlation of 0.8 and 10 items is not sufficient to increase group-level $\alpha$ above 0.7.
Cheng et al. (2012) **Estimator(s):** IRT-based coefficient ($\pi$) (Samejima, 1994) Samejima, $\omega$, maximum likelihood FA coefficient (Bentler, 1968) • **Data generation:** Data generation procedure described, IRT unidimensional graded response model, no software reported, no data validity check, no code provided • **Replications:** 1000 / 10 / not reported, numbers not justified • **Conditions:** Simullee population, 30 items, IRT parameters following intervals from uniform distribution: $a$ (0.5, 1.5), $b$ (−3, 3) / simullee population, 30 items, IRT parameters following intervals from uniform distribution: $a$ (0, 1.5), $b$ (−3, 3), number of scale points (0−15) / simullee population, 30 items, IRT parameters following intervals from uniform distribution: $a$ (0, 2.5), $b$ (−3, 3), number of scale points (0−15), selection of conditions partially justified (in the first study)—$a$ and $b$ parameter levels are chosen to reflect those of typical psychological and educational assessments • **Benchmark:** – • **Performance evaluation:** Values across replications, differences between estimator values, visualization used • **Conclusion(s):** $\omega$ yields higher values than $\pi$ typically; as the number of scale points increases, $\pi$ can surpass $\omega$.

Cui and Li (2012) **Estimator(s):** 6 CI estimators of $\alpha$ based on parametric approach (e.g., Feldt, 1965; Hakstian & Whalen, 1976) and bootstrapping • **Data generation:** Data generation procedure described, RNG, Mathematica, no data validity check, code available upon request • **Replications:** 1000, number not justified • **Conditions:** Simullee sample size (20, 50, 100), number of items (5, 20, 40), population reliability (0.4, 0.6, 0.8), measurement model type (parallel, tau-equivalent, congeneric), distribution (normal, non-normal), number of scale points (continuous, 2, 5), selection of conditions justified—factors and levels justified • **Benchmark:** Population reliability / $\alpha$ • **Performance evaluation:** Empirical coverage, CI width, visualization not used • **Conclusion(s):** Performances of bootstrap percentile interval (BPI) and bootstrap bias corrected percentile interval (BCaI) are not satisfactory; bootstrap standard interval (BSI) and parametric CIs perform well when distribution is normal.

Estabrook et al. (2012) **Estimator(s):** Intraindividual SD (ISD), intraindividual M (IIM) • **Data generation:** Data generation procedure described, time series model, R, no data validity check, code available upon request with sample code in Appendix • **Replications:** 100, number not justified • **Conditions:** 100 simulees, population IIM (0.2, 0.4, ..., 2), population ISD (0.08, 0.16, 0.24, 0.32, 0.4, 0.48), number of time points (3, 4, ..., 9, 10, 15, 20, 25, 30, ..., 95, 100), population reliability / internal consistency at each time point (0.6, 0.7, 0.8, 0.9), selection of conditions partially justified—factor levels justified • **Benchmark:** Population reliability specified for each time point • **Performance evaluation:** Descriptive statistics (median), regression analysis—Fischer’s $z$-transformed reliability regressed on linear and natural logarithm transformations of the four factor parameters and their interactions, visualization used • **Conclusion(s):** ISD has low reliability; researchers should check if floor or ceiling effect is present, whereas IIM provides partial information about it.

Fife et al. (2012) **Estimator(s):** K-R 20 coefficient, $\omega$, coefficient of stability • **Data generation:** Data generation procedure described, IRT model, no software reported, no data validity check, no code provided • **Replications:** 10000, number not justified • **Conditions:** 20 dichotomous items, 500 simulees, three IRT $a$ (0.45, 0.65, 1.69) parameters, population reliability (0.7, 0.8, 0.94), amount of range restriction (0.1, 0.2, ..., 1) / 20 dichotomous items, 500 simulees, three IRT $a$ (0.45, 0.65, 1.69) parameters, population reliability (0.7, 0.8, 0.94), amount of range restriction (0.1, 0.2, ..., 1), selection
of conditions partially justified—factor levels justified • Benchmark: Global (based on whole population) and local (based on a subset of population consisting of 10% top performers) reliability based on 10 000 simulees • Performance evaluation: Bias (relative bias), precision (SD of estimator values) / SE of estimator values compared to first simulation, visualization used • Conclusion(s): Coefficient of stability is the best estimator of reliability across a variety of conditions; K-R 20 and $\omega$ perform well under indirect range restriction; all estimators are more imprecise as the range restriction increases.

Padilla et al. (2012) Estimator(s): 3 bootstrap $\alpha$ CI estimators (normal theory, percentile, bias-corrected and accelerated), 3 non-bootstrap $\alpha$ CI estimators (Bonett, 2003; Fisher, 1950; Maydeu-Olivares et al., 2007) • Data generation: Data generation procedure described, RNG, no software reported, data validity partially checked—values of population reliability using relative bias, no code provided • Replications: 1000 for each bootstrap estimator, 25 000 for each non-bootstrap estimator, numbers not justified • Conditions: Number of items (5, 10, 15, 20), item correlation matrix (based on parallel measurement model with factor loadings of 0.55 or 0.745, based on congeneric measurement model with loadings of 0.3, 0.4, ..., 0.7), number of scale points (2, 3, 5, 7), distribution shape (type 1: skewness = 0, kurtosis = -2, type 2: skewness = 1.7, kurtosis = 0.88, type 3: skewness = 0.41, kurtosis = -1.83), simulee sample size (50, 100, 150, 200, 250, 300), selection of conditions partially justified—factor levels justified • Benchmark: Population reliability / $\alpha$ specified for each number of items and number of scale points condition • Performance evaluation: Coverage, visualization used • Conclusion(s): The normal theory bootstrap method had the best performance by far, the methods proposed by Bonett (2003) followed by the normal theory method were good alternatives, the Fisher method performed unsatisfactorily under most conditions.

Sheng and Sheng (2012) Estimator(s): $\alpha$ • Data generation: Data generation procedure described, RNG, MATLAB, no data validity check, code provided in Appendix • Replications: 100 000, number not justified • Conditions: Tau-equivalent measurement model, simulee sample size (30, 50, 100, 1000), number of items (5, 10, 30), population reliability (0.3, 0.6, 0.8), distribution (normal, symmetric platykurtic, symmetric leptokurtic, non-symmetric, non-symmetric platykurtic, non-symmetric leptokurtic), target of transformation to non-normality (true scores, errors), selection of conditions partially justified—population reliability levels • Benchmark: Population reliability with normal distribution for each condition • Performance evaluation: Descriptive statistics (M, precision (SD), CI), bias, RMSE, no visualization used • Conclusion(s): The sample $\alpha$ is affected by leptokurtic true score distribution, or skewed and/or kurtotic error score distribution; increased sample sizes, not test lengths, help improve the accuracy, bias, or precision with non-normal data.

Wang and Grimm (2012) Estimator(s): Intraindividual SD (ISD), intraindividual variance (ISD$^2$) intraindividual M (IIM) • Data generation: Data generation procedure described, proposed detrended time series model (Wang & Grimm, 2012) for modeling scores within a burst from a measurement-burst design or intensive longitudinal design, R, no data validity check, no code provided • Replications: 1 000 000, number not justified • Conditions: 1 000 000 simulees, distribution of heterogeneous intraindividual variability (6 gamma distribution shapes), amount of measurement error variance (0.05,
Monte Carlo simulation studies of reliability in psychometrics

0.125, 0.25, 0.5, 1, 2), population reliability (0.3, 0.5, 0.7, 0.9), number of assessments per burst (5, 25) / 1 000 000 simulees, distribution of heterogeneous intraindividual variability (6 gamma distribution shapes), amount of measurement error variance (0.05, 0.125, 0.25, 0.5, 1, 2), population reliability (0.3, 0.5, 0.7, 0.9), number of assessments per burst (3–150), selection of conditions not justified • Benchmark: Population reliability • Performance evaluation: Correlation between estimator values and theoretical values / descriptive statistics (M) compared to population reliability, pairwise differences of estimators per condition, ratio of inter-individual SD of Ms and inter-individual SD of ISDs, visualization used • Conclusion(s): The reliability of IIM under every condition and the reliability of ISD is lower than that of ISD under most conditions.

F. Gu et al. (2013) Estimator(s): $\alpha$, SEM-based coefficient (Raykov, 1997) • Data generation: Data generation procedure described, RNG, SAS, no data validity check, code available online • Replications: 2000, number not justified • Conditions: 1000 simulees, 15 items, 3 dimensions with 5 items in each, number of items violating tau-equivalence equally divided across dimensions (3, 6), amount of true-score shrinkage for violating items (1, 0.9, 0.7, 0.6, 0.3), ratio of true to error variance (1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1), error correlatedness (0, 0.1, 0.2, 0.3, 0.4), selection of conditions not justified • Benchmark: Population reliability • Performance evaluation: Descriptive statistics (M, SD) compared to population reliability, bias, visualization not used • Conclusion(s): Both coefficients are unbiased if assumptions are fulfilled; the bias of $\alpha$ is more severe than the bias of SEM-based coefficient in case of extreme violation of assumptions.

Headrick and Sheng (2013) Estimator(s): $\alpha$, variant of $\alpha$ based on L-comoments (Headrick & Sheng, 2013) • Data generation: Data generation procedure described, RNG, MATLAB, no data validity check, no code provided • Replications: 25 000, number not justified • Conditions: Simulee sample size (10, 20, 1000), number of items (4, 9, 10), true score distribution (standard normal, symmetrical leptokurtic, asymmetrical leptokurtic), ratio of diagonal to off-diagonal values (2, 5), method of obtaining item variance-covariance matrices (conventional, L-comoments) selection of conditions not justified • Benchmark: Population reliability / $\alpha$ specified for each true score distribution, ratio of diagonal to off-diagonal values, and number of items condition • Performance evaluation: Descriptive statistics (bias-corrected accelerated bootstrapped M, CI, SE, relative SE), relative bias, visualization not used • Conclusion(s): $\alpha$ based on L-comoments (Headrick & Sheng, 2013) is useful when distribution is heavy-tailed and sample size is small.

Padilla and Divers (2013) Estimator(s): 3 bootstrap CI estimators for $\omega$ coefficient (normal theory, percentile, bias-corrected and accelerated Bonett, 2003; Fisher, 1950; Maydeu-Olivares et al., 2007) • Data generation: Data generation procedure described, RNG, no software reported, data validity partially checked—relative bias of point estimates, no code provided • Replications: 1000, number not justified • Conditions: Number of items (5, 10, 15, 20), item correlation matrix (based on parallel measurement model with factor loadings of 0.55 or 0.745, based on congeneric measurement model with loadings of 0.3, 0.4, ..., 0.7), number of response categories (2, 22, 3, 5, 7), simulee sample size (50, 100, 150, 200), selection of conditions partially justified—factor levels justified • Benchmark: Population reliability / $\omega$ specified for each number of items and number of scale points condition • Performance evaluation: Coverage, visualization
used. Conclusion(s): All methods performed well; the normal theory bootstrap CIs outperformed other methods with more consistent acceptable coverage.

Şimşek and Noyan (2013) Estimator(s): $\alpha$, $\omega$, generalized $\theta$ • Data generation: Data generation procedure described, RNG, SAS, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: 6 items, factor structure (2 dimensions, tau-equivalent and congeneric factors), correlation between dimensions ($-0.9$, $-0.7$, ..., 0.7, 0.9), simulee sample size (50, 100, 200, 400), selection of conditions not justified • Benchmark: Population reliability specified for each coefficient under each condition of factor structure and correlation between dimensions • Performance evaluation: Bias, relative bias, efficiency (SD of replicated values), precision (RMSE), visualization not used • Conclusion(s): Generalized $\theta$ outperformed $\alpha$ and $\omega$, especially with orthogonal or nearly orthogonal dimensions.

Geldhof et al. (2014) Estimator(s): $\alpha$, $\omega$, maximal reliability (H; Thomson, 1940), multilevel reliability coefficient (Raykov & Marcoulides, 2006) • Data generation: Data generation procedure described, RNG, Mplus, data validity partially checked—model fit was examined, no code provided • Replications: 1000, number justified—as an arbitrary balance between generating a large enough number of replications to obtain appropriate precision of our estimates and the time required to analyze, and because that number was used in previous studies • Conditions: 6 items, congeneric model, simulee sample size (2, 15, 30), number of clusters (50, 100, 200), item intraclass correlation (0.05, 0.25, 0.5, 0.75), level(s) of analysis with high reliability / population reliability (within only, between only, both levels, neither level), levels with high reliability $\lambda_1 = \lambda_2 = 0.8$, $\lambda_3 = \lambda_4 = 0.7$, $\lambda_5 = \lambda_6 = 0.6$, $\alpha = 0.852$, $\omega = 0.854$, $H = 0.372$, levels without high reliability $\lambda_j = 0.3$, $\alpha = \omega = H = 0.372$, selection of conditions not justified • Benchmark: Population reliability specified for each condition • Performance evaluation: Descriptive statistics (CI, Bayesian credible intervals), percent bias at each level, ANOVA with $\eta^2$, visualization used • Conclusion(s): Single-level estimates do not correspond to population reliability unless reliability is identical at each level of analysis; 2-level $\alpha$ and $\omega$ perform relatively well under most conditions; estimates of maximal reliability (H) are more biased in case of multilevel data than $\alpha$ and $\omega$; small cluster size may cause positive bias of reliability in case of between-level analysis.

Cuesta Izquierdo and Fonseca-Pedrero (2014) Estimator(s): $\alpha$ • Data generation: Data generation procedure described—based on an empirical dataset (Fonseca-Pedrero et al., 2010), RNG, SPSS, no data validity check, no code provided • Replications: 100, number not justified • Conditions: 200 simulees, 51 items, 5 scale points, missing data mechanism (2 to 1 or 3 to 1 in favor of one group), percentage of missing values (5, 10, 20, 30), simulee sample size (3056, 200 random subsamples), missing data treatment procedure (listwise, imputation by multiple regression, imputation by expectation maximization (EM) procedure, imputation by “simple EM procedure”, multiple imputation by a sequential regression procedure), selection of conditions not justified • Benchmark: Population reliability, based on value of $\alpha$ in the empirical data of a questionnaire • Performance evaluation: Descriptive statistics (M, minimum, maximum), RMSE, average bias, discrepancy—average difference in $\alpha$ between the complete data and data with treated missing values, regression analysis, ANOVA with RMSE, bias, and discrepancy as DVs and $\eta^2$ • Conclusion(s): With a small percentage of missing values it is possible to obtain acceptable estimations from a practical point
of view with all the procedures employed, except listwise.

Shu and Schwarz (2014) *Estimator(s)*: IRT variants of $\alpha$, Feldt’s (1975) and Raju’s (1977) coefficients, stratified $\alpha$ (Cronbach et al., 1965) and marginal reliability • *Data generation*: Data generation procedure described, 3-PL and 2-PL partial credit IRT models, no software reported, data validity partially checked—model fit was examined, no code provided • *Replications*: Not reported • *Conditions*: 2000 simulees, 36 items (9 closed and 27 open; 9 with 2, 3 and 4 levels), IRT ability distribution $[N(0, 1), N(-0.16, 1.18), N(-0.5, 1.25), N(1.5, 1)]$, selection of conditions not justified • *Benchmark*: Population reliability specified for each condition • *Performance evaluation*: Descriptive statistics (M) compared to population reliability, visualization not used • *Conclusion(s)*: IRT variants of internal consistency coefficients outperform CTT coefficients.

Bonanomi et al. (2015) *Estimator(s)*: Ordinal $\alpha$ coefficient based on polychoric correlation (Zumbo et al., 2007), ordinal $\alpha$ coefficient based on copula measures of association instead of polychoric correlation (Bonanomi et al., 2015) • *Data generation*: Data generation procedure described, RNG, no software reported, no data validity check, no code provided • *Replications*: 1000, number not justified • *Conditions*: 1000 simulees, number of scale points (3, 7), Spearman correlation between copulae (0.3, 0.7, 0.9), selection of conditions not justified • *Benchmark*: Population reliability / theoretical $\alpha$ • *Performance evaluation*: Descriptive statistics (M), bias, MSE, visualization not used • *Conclusion(s)*: Ordinal $\alpha$ coefficient based on copula measures of association (Bonanomi et al., 2015) is more robust to the distributional assumption, showing a lower MSE.

Hunt and Bentler (2015) *Estimator(s)*: Cumulative $\lambda_4(0.05,0.5,0.95)$ coefficients, GLB (Moltenner & Revelle, 2016), $\alpha$ • *Data generation*: Data generation procedure described, FA model, R, no data validity check, no code provided • *Replications*: 500, number not justified • *Conditions*: 16 items, simulee sample size (50, 100, 400, 1000, 2000), one dimension, tau-equivalent items with loadings 0.6 and error variances 0.6, 0.7, 0.8, 0.9 repeated 4 times / 16 items, simulee sample size (50, 100, 400, 1000, 2000), two dimensions correlated by 0.3, items with loadings 0.6 and error variances 0.6, 0.7, 0.8, 0.9, repeated twice for both factors / 16 items, simulee sample size (50, 100, 400, 1000, 2000), two dimensions correlated by 0.3, items with loadings 0.7, 0.8, 0.9 repeated across 16 items and error variances 0.6, 0.7, 0.8, 0.9, selection of conditions not justified • *Benchmark*: Population reliability specified for each simulation • *Performance evaluation*: Descriptive statistics (M, SD), bias, visualization used • *Conclusion(s)*: $\alpha$ displays negative bias, whereas GLB displays positive bias under nearly every condition; $\lambda_4(0.05)$ is the least prone to positive bias, $\lambda_4(0.50)$ is the median of these coefficients, while $\lambda_4(0.95)$ is inclined to positive bias, but less than Guttman’s maximal $\lambda_4$.

Paek (2015) *Estimator(s)*: $\alpha$, IRT-based coefficient • *Data generation*: Data generation procedure described, 3-PL IRT model, no software reported, no data validity check, no code provided • *Replications*: 100, number not justified • *Conditions*: Number of items (20, 40, 60), item discrimination (0.6, 0.9, 1.2), item difficulty ($-1.5, 0, 1.5$), guessing parameter (0.08, 0.25, 0.42), selection of conditions partially justified—selection of factors justified as to make the conditions more realistic • *Benchmark*: – • *Performance evaluation*: Descriptive statistics (M), ANOVA with $\eta^2$, correlation between estimators, visualization used • *Conclusion(s)*: In addition to the general negative impact of
guessing on reliability, results showed interaction effects between number of items and guessing and between guessing and difficulty.

Hu et al. (2016) *Estimator(s)*: Spearman-Brown formula as coefficient of equivalence, SEM-based reliability coefficient (Raykov & Marcoulides, 2015), intraindividual variance

- **Data generation**: Data generation procedure described, RNG / autoregressive integrated moving average (ARIMA) model, R functions *rnorm()* and *arima.sim()* , no data validity check, no code provided • **Replications**: Not reported • **Conditions**: 3 items, 5 simulees, 100 time points, conditions (only unique variances differing across simulees, only factor loadings differing across simulees, unique variances and factor loadings invariant across simulees, violated independence assumption for one simulee), selection of conditions not justified • **Benchmark**: Population reliability / SEM-based coefficient • **Performance evaluation**: Descriptive statistics (M, CI) compared to population reliability, visualization used • **Conclusion(s)**: Results support the use of SEM and parallel forms to estimate reliability, as well as justify estimating reliability at the individual level using repeated measures.

Kelley and Pornprasertmanit (2016) *Estimator(s)*: CI estimators of $\alpha$, $\omega$, $\omega_h$ and $\omega_{\text{categorical}}$ using approach of Feldt (1965), delta method (Maydeu-Olivares et al., 2007), transformation based approaches (Fisher, 1950; Hakstian & Whalen, 1976), likelihood-based approach and bootstrap approaches, 13 methods for $\alpha$ and 15 for $\omega$ overall, in the first study $\alpha$ and $\omega$ CI estimators, in second study $\omega_h$ and $\omega$ CI estimators, in the third study $\omega_h$ and $\omega_{\text{categorical}}$ CI estimators • **Data generation**: Data generation procedure described, RNG, R function *mvrnorm()* and *simulateData()* function from package *lavaan*, no data validity check, no code provided • **Replications**: 1000, number not justified • **Conditions**: Simulee sample size (50, 100, 200, 400, 1000), number of items (4, 8, 12, 16, 20), factor loading distribution (tau-equivalent, non tau-equivalent), population reliability (0.7, 0.8, 0.9), item distribution (multivariate normal, skewness/kurtosis 1.25/3.5, 2.25/7, 3.25/20) / simulee sample size (50, 100, 200, 400, 1000), number of items (4, 8, 12), population reliability / $\omega$ (0.7, 0.8, 0.9), model error (RMSEA of 0.02, 0.05, 0.08, 0.1) / congeneric measurement model, simulee sample size (50, 100, 200, 400, 1000), number of items (4, 8, 12), number of scale points (2, 5), threshold symmetry (symmetry, moderate asymmetry, extreme asymmetry, moderate asymmetry-alternating, extreme asymmetry-alternating), population reliability / $\omega_{\text{categorical}}$ (0.7, 0.8, 0.9), selection of conditions partially justified—factor levels justified • **Benchmark**: Population reliability / $\omega$ / $\omega_{\text{categorical}}$ • **Performance evaluation**: Coverage, ANOVA with coverage rate as DV and $\eta^2$, visualization not used • **Conclusion(s)**: Bootstrap CIs for hierarchical omega for continuous items and categorical omega for categorical items are recommended.

Padilla and Divers (2016) *Estimator(s)*: 3 non-bootstrap CI estimators of $\omega$ (normal theory, delta with logit transformation method, the three-step parceling method), 4 bootstrap CI estimators of $\omega$ (empirical sampling distribution, percentile-based, 2 methods proposed by Padilla & Divers, 2013 • **Data generation**: Data generation procedure described, RNG, no software reported, no data validity check, no code provided • **Replications**: 1000, number not justified • **Conditions**: Simulee sample size (50, 100, 150, ..., 500), number of items (6, 12, 18, 24), measurement model type (3 congeneric models with loading structures: [0.4, 0.5, ..., 0.9], [0.3, 0.4, ..., 0.8], [0.4, 0.4, 0.4, 0.8, 0.8, 0.8]), number of scale points (2, 3, 5, 7), distribution (type 1: skewness = 0, kurtosis = −2,
type 2: skewness = 1.7, kurtosis = 0.88, type 3: skewness = 0.41, kurtosis = −1.83), selection of conditions partially justified—number of items to be comparable with previous research • Benchmark: Population reliability / ω for each number of items and number of scale points condition • Performance evaluation: Coverage, visualization used • Conclusion(s): The normal theory bootstrap CI estimator outperformed other estimators under every condition with sample size below 100; most methods had sound coverage with sample sizes of 100 or more.

Stanley and Edwards (2016) Estimator(s): α, ω • Data generation: Data generation procedure described—based on an empirical dataset—PROMIS Anxiety Item Bank (Pilkonis et al., 2011), IRT graded response model, flexMIRT software package, no data validity check, no code provided • Replications: 100, number not justified • Conditions: 1000 simulees, 10 items, item slopes multiplied so as to represent an increasingly unreliably measured construct (1, 1/2, 1/3, 1/4, 1/5, 1/6), selection of conditions not justified • Benchmark: Population reliability, based on coefficient values in the empirical data of a questionnaire • Performance evaluation: Average bias, RMSE • Conclusion(s): Reliability and model fit are distinct and disagreement between reliability and model fit may provide information on model validity.

Trizano-Hermosilla and Alvarado (2016) Estimator(s): α, ω, GLB (R. W. B. Jackson & Ferguson, 1941), GLBalg (Moltner & Revelle, 2016) • Data generation: Data generation procedure described, FA model, R, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: Simulee sample size (250, 500, 1000), number of items (6, 12), measurement model type (tau-equivalent, congeneric), progressive incorporation of asymmetrical items (0 % to 100 %), selection of conditions not justified • Benchmark: Population reliability specified for each number of items condition • Performance evaluation: Descriptive statistics (M, SD), RMSE, percent bias (relative bias), visualization not used • Conclusion(s): When all items are normally distributed, ω should be used, in case of low to moderate skewness GLBalg should be used, in case of high proportion of asymmetrical items, GLB should be used.

Zhang and Yuan (2016) Estimator(s): Robust variants of α and ω point and CI estimators • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: 100 simulees, 6 items, measurement model type (tau-equivalent, congeneric), data type (normally distributed, 5 % outliers, 5 % leverage observations) / 100 simulees, 6 items, 6.7 % missing data with 20 % to 40 % of simulees having at least 1 missing value, measurement model type (tau-equivalent, congeneric), method of handling missingess (listwise deletion, robust estimators), selection of conditions not justified • Benchmark: Population reliability / α and ω for each measurement model type condition (identical under tau-equivalence) • Performance evaluation: Descriptive statistics (M, empirical SE based on bootstrapping), coverage, visualization not used • Conclusion(s): For normally distributed data, there is no substantial difference between robust variants of α and ω regardless if tau-equivalence holds; outliers cause underestimation of alpha and omega whereas leverage observations cause their overestimation; robust variants of α and ω are more useful than listwise deletion under conditions with missing data.

Turner et al. (2017) Estimator(s): 4 CI estimators of ordinal α (Bonett, 2002; Feldt, 1965; Fisher, 1950; Hakstian & Whalen, 1976) • Data generation: Data generation proce-
dure described, FA model, R, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: Population reliability (0.6, 0.8, 0.9), simulee sample size (20, 50, 100, 200), number of items (5, 10, 25, 40), skewness (0, −1.217), number of scale points (5, 7), selection of conditions partially justified—factor levels justified based on previous studies and that are often noted in applied research • Benchmark: Population reliability / α for each condition based on 1 000 000 simulees • Performance evaluation: Coverage, precision (RMSE), ANOVA with RMSE, bias, CI width, and variance of CI width as DVs and η^2, visualization used • Conclusion(s): All methods yielded unacceptably low coverage rates and potentially increased type I error rates.

Andersson and Xin (2018) Estimator(s): Test and marginal IRT reliability • Data generation: Data generation procedure described, generalized partial credit model and 3-PL IRT models, R, no data validity check, no code provided • Replications: 2000, number not justified • Conditions: Simulee sample size (250, 500, 1000, 2000, 4000, 8000 for 3-PL model, and 1000, 2000, 4000, 8000 for generalized partial credit model), selection of conditions partially justified—simulee sample size levels justified for the model convergence rate • Benchmark: Population reliability for test and marginal reliability for each IRT model separately • Performance evaluation: Mean bias, asymptotic SE, Monte Carlo SE, coverage, visualization not used • Conclusion(s): Confidence intervals for the test reliability coefficient typically have good coverage properties in finite samples under with the generalized partial credit and 3-PL models; estimator of the marginal reliability coefficient has finite sample bias resulting in confidence intervals that do not attain the nominal level for small sample sizes but that the bias tends to zero as the sample size increases.

Du and Wang (2018) Estimator(s): Intraindividual SD (ISD), intraindividual variance (ISD^2), h-th order autocorrelation coefficient (c), mean square successive difference (MSSD) • Data generation: Data generation procedure described, multilevel time series model, no software reported, no data validity check, no code provided • Replications: 1 000 000, number not justified • Conditions: 1 000 000 simulees, scale population reliability (0.3, 0.5, 0.7, 0.9, 1), variation in trait score (0.1, 5, 10), ratio of average true intraindividual variation to interindividual variation in trait scores (0.05, 0.125, 0.25, 0.5, 1, 2), ratio of average true intraindividual variation to interindividual variation in variances (0.05, 0.125, 0.25, 0.5, 1, 2), average autocorrelation parameter (0.1, 0.3, 0.5), SD of individual autocorrelation coefficients, number of assessments (3, 5, 10, 15, 20, 30, 40, ..., 100, 150, 200, 300, 400, 500), selection of conditions justified—to cover most conditions in behavioral and psychological data • Benchmark: Population reliability • Performance evaluation: Descriptive statistics (M of SE), pairwise differences, visualization used • Conclusion(s): Indicators of intraindividual variability were more reliable with a more reliable measurement scale and more assessments; ISD and ISD^2 were the most reliable, followed by MSSD and r_{(h)}.

Trinchera et al. (2018) Estimator(s): 5 CI estimators of α (Bonett & Wright, 2015b; Feldt, 1965; Raykov & Marcoulides, 2015; Trinchera et al., 2018) • Data generation: Data generation procedure described, FA model as in Raykov and Marcoulides (2015), no software reported, no data validity check, no code provided • Replications: 10 000, number not justified • Conditions: Congeneric measurement model (loadings specified for each number of items condition), number of items (3, 5, 7, 9), simulee sample size
Monte Carlo simulation studies of reliability in psychometrics

(30, 50, 100, 150, 250), selection of conditions partially justified—number of item levels selected based on previous studies • Benchmark: Population reliability / $\alpha$ • Performance evaluation: Descriptive statistics (M), coverage, visualization used • Conclusion(s): CI estimator (Trinchera et al., 2018) yields smaller SE compared to other estimators, especially for small sample sizes.

Zijlmans, van der Ark, et al. (2018) Estimator(s): Molenaar and Sijtsma’s (MS; 1988) method, latent class reliability coefficient (LCRC; Van Der Ark et al., 2011), Zijlmans, Tijmstra, et al.’s (2018), Wanous and Reichers (1996) method (CA) for item score reliability estimation, Guttman’s $\lambda_6$ • Data generation: Data generation procedure described, multidimensional graded IRT model, R, no data validity check, code provided online • Replications: 1000, number not justified • Conditions: Simulee sample size (200, 1000), number of items (6, 18) two dimensions correlated by 0.5, item type (dichotomous/polytomous items with 5 levels), two IRT a parameters (0.5, 2), selection of conditions partially justified—conditions similar as in Van Der Ark et al. (2011), but dimensionality specified to be more realistic • Benchmark: Population reliability based on scores for 1000 simulee with 6 dichotomous items, one dimension, equal discrimination parameters and equidistantly spaced location parameters • Performance evaluation: Median bias, interquartile range, percentage of outliers, visualization used • Conclusion(s): MS and CA are the least biased; LCRC displays nearly unbiased values, but high variability; $\lambda_6$ displays consistent negative bias, but lowest variability.

S.-J. Cho et al. (2019) Estimator(s): 2 multilevel reliability coefficients based on estimators derived by S.-J. Cho et al. (2015), Monte Carlo confidence and Bayesian credible intervals • Data generation: Data generation procedure described, 2-PL multilevel IRT model with either separate or equal discrimination across levels, Mplus, data validity partially checked—model convergence was examined and results were cross-validated using Mplus, flexMIRT and glm() function of SAS, code available upon request • Replications: 200, number not justified • Conditions: Number of items (20, 40), number of multilevel clusters (20, 40, 100, 200), cluster size (15, 30), item-specific intraclass correlation (0.1, 0.2) / number of items (20, 40), number of multilevel clusters (20, 40, 100, 200), cluster size (15, 30), item-specific intraclass correlation 0.5, selection of conditions justified—factor selection and levels justified based on previous studies • Benchmark: Population multilevel reliability (within and between levels) • Performance evaluation: Accuracy (relative percentage bias, RMSE), coverage rates estimated by marginal maximum likelihood (MMLE)-multiple imputation and Bayesian analysis, visualization used • Conclusion(s): MMLE-multiple imputation is generally recommended.

Yang and Xia (2019) Estimator(s): Categorical $\omega$ (nonlinear SEM coefficient by Green & Yang, 2009) computed within Bayesian framework • Data generation: Data generation procedure described, FA model, Mplus, data validity partially checked—number of iterations was set to meet the convergence criterion, no code provided • Replications: 2000, number not justified • Conditions: Factor structure (unidimensional structure with no correlated residuals, bifactor structure with all items loaded on a general factor and subgroups of items loaded on 2 or 4 group factors), number of items (8, 16), number of scale points (2, 4), scale heterogeneity (based on three groups of thresholds for categorization), simulee sample size (50, 100, 200, 500), 5 prior sets obtained by combining the priors for loadings and priors for thresholds, selection of conditions
partially justified—factor structure selection justified and it received much attention in the literature • Benchmark: Population categorical \( \omega \) for each factor structure and number of items conditions • Performance evaluation: Descriptive statistics (M, SD), absolute bias, predictive posterior \( p \)-value, visualization used • Conclusion(s): Bayes estimator of categorical \( \omega \) is a promising method, although it performs differently depending on the conditions.

Gurdil Ege and Demir (2020) Estimator(s): \( \alpha \), stratified \( \alpha \), Angoff-Feldt coefficient (Feldt, 1975), Feldt-Raju coefficient (Raju, 1977) • Data generation: Data generation procedure described, partial credit and 2-PL IRT model, WinGen, no data validity check, no code provided • Replications: 25, number justified based on previous study and literature (Harwell et al., 1996) • Conditions: 30 items, simulee sample size (500, 1000, 2000), number of scale points (2, 4), ratio of dichotomous to polytomous items (1:1, 1:2, 2:1), selection of conditions partially justified—simulee sample size and number of scale points levels justified based on literature • Benchmark: Population reliability • Performance evaluation: Descriptive statistics (M, SE of M), ANOVA with \( \eta^2 \), visualization not used • Conclusion(s): Angoff-Feldt, and Feldt-Raju reliability coefficients were higher when the number of dichotomous items in the item-type ratio was higher than that of polytomous items, while the reverse was found for \( \alpha \) and stratified \( \alpha \); estimators give similar results overall when sample size is at least 1000.

Hancock and An (2020) Estimator(s): \( \omega \) based on confirmatory FA and closed form of \( \omega \) • Data generation: Data generation procedure described, FA model, package lavaan in R, no data validity check, no code provided • Replications: 100, number not justified • Conditions: Number of items (3, 6, 9, 12), factor loading magnitude/variability (low/low (0.3, 0.4, 0.5; \( \omega = 0.365 \)), low/high (0.1, 0.4, 0.7; \( \omega = 0.381 \)), high/low (0.5, 0.6, 0.7; \( \omega = 0.63 \)), high/high (0.3, 0.6, 0.9; \( \omega = 0.651 \)), simulee sample size (250, 500, 1000, 5000), item scale (normal continuous, dichotomous, uniform 5 point, normal 5 points), selection of conditions not justified • Benchmark: Population reliability / \( \omega \) for each condition based on scores of 1 000 000 simulees • Performance evaluation: Descriptive statistics (precision (SD), relative precision (ratio of SD of estimator values and SD of generated data under each condition), estimate failure rate, relative percent bias, visualization not used • Conclusion(s): Closed form of \( \omega \) may be used instead of \( \omega \) as a more accessible computational alternative for practitioners.

Kim et al. (2020) Estimator(s): \( \alpha \), nonlinear SEM-based coefficient (Kim et al., 2020) • Data generation: Data generation procedure described, FA model, no software reported, data validity partially checked—model convergence was examined, no code provided • Replications: 100, number not justified • Conditions: 500 simulees, number of items (9, 18), factor model (one-factor, bi-factor with 3 group factors with 3 or 6 items per factor), factor loadings (0.7 for general factor, 0.4 or 0.6 for group factors), number of scale points (2, 5), distribution of thresholds for categorization (normal, moderate skew, mixed skew), selection of conditions partially justified—number of simulees justified to avoid estimation problems due to small sample size as sample size was not the focus of the study • Benchmark: Population reliability as the correlation between parallel forms based on two populations of 100 000 simulees • Performance evaluation: Descriptive statistics (M, SD) compared to population reliability, visualization not used • Conclusion(s): \( \alpha \) is close to population reliability under conditions of one factor and equal number of scale points; proposed SEM-based reliability coefficient is
close to population reliability in most conditions.

Edwards et al. (2021) *Estimator(s): α, ω, ω_{hierarchical}, Revelle’s ω, GLB* • *Data generation:* Data generation procedure described, FA model, R function `simulateData()` from package `lavaan`, no data validity check, no code provided but output is in the Appendix • *Replications:* 1000, number not justified • *Conditions:* Measurement model type (three parallel conditions and 11 tau-equivalent taken over from Green & Yang, 2009), simulee sample size (30, 50, 100, 500, 1000), number of items (6, 12), selection of conditions partially justified—common conditions observed in psychology and educational sciences that may impact internal consistency • *Benchmark:* Population reliability based on each measurement model and number of items condition • *Performance evaluation:* Descriptive statistics (M, SE of coefficient values) compared to population reliability, regression analysis using α and ω as DVs, visualization used • *Conclusion(s):* α and ω yielded the most accurate reflection of population reliability values.

Fu et al. (2021) *Estimator(s): ω based on confirmatory FA (ICM-CFA) and ω based on exploratory SEM (ESEM) with geomin rotation* • *Data generation:* Data generation procedure described, FA model, Mplus, data validity partially checked—model fit of proper solutions examined, no code provided • *Replications:* 500, number not justified • *Conditions:* Two dimensions correlated by 0.3, number of items per factor (3, 6 with 2 or 4 cross-loadings), factor loading (0.4, 0.6, 0.8), cross-loading value (0.01, 0.1, 0.25), simulee sample size (100, 300, 500) / number of items per factor (3, 6), cross-loading value (0.01, 0.1, 0.25), simulee sample size (100, 300), selection of conditions partially justified—simulee sample size levels justified based on literature • *Benchmark:* Population reliability • *Performance evaluation:* Relative bias, SE, MSE, visualization not used • *Conclusion(s):* If the model fit of ESEM is acceptable but that of ICM-CFA is not, the composite reliability estimates based on the above two models should be similar: if the target factor loadings are relatively small, researchers should increase the number of items per factor or increase the sample size.

Garcia-Garzon et al. (2021) *Estimator(s): ω_{h} with 6 algorithms (bi-quartimin, bi-geomin, Schmid-Leiman (SL), empirical iterative empirical target rotation based on an initial SL solution (SLiD), direct SL (DSL), and direct bi-factor (DBF)* • *Data generation:* Data generation procedure described, RNG, R, no data validity partially checked—convergence was examined, no code provided • *Replications:* 100, number not justified • *Conditions:* Simulee sample size (150, 500, 1000), number of group factors (3, 6), number of items per factor (4, 8), cross-loadings on the group factors (yes, no), pure indicators of the general factor (yes, no), general factor loading (0.3, 0.45, 0.6) / simulee sample size (150, 500, 1000), number of group factors (3, 6), number of factor indicators (4, 8), second-order factor loading (0.55, 0.7, 0.8) / simulee sample size (150, 500, 1000), number of group factors (3, 6), number of factor indicators (4, 8), cross-loadings on the group factors (yes, no), selection of conditions not justified • *Benchmark:* Population reliability / ω_{h} specified for each condition involving factor structure • *Performance evaluation:* Accuracy (mean absolute error (MAE) and mean bias error (MBE)), RMSE, ANOVA with MAE and MBE as DVs with partial η^2, visualization used • *Conclusion(s):* SLiD provided the best approximation to ω_{hierarchical} under most conditions; neither SL, bi-quartimin, nor bi-geomin produced an overall satisfactory recovery of omega hierarchical; the performance of DSL and DBF depended upon the average discrepancy between the loadings of the general and the group factors.
Pfadt et al. (2022) Estimator(s): $\alpha$, Guttman’s $\lambda_2$, $\omega$, GLB within frequentist and Bayesian framework • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: 1000, number not justified • Conditions: Average inter-item correlation (0, 0.3, 0.7), number of items (5, 20), simulee sample size (50, 100, 500), selection of conditions partially justified—average inter-item correlation levels justified • Benchmark: Population reliability specified for each coefficient • Performance evaluation: Descriptive statistics (M, SE), RMSE, coverage, risk (probability of overestimating the population value), visualization used • Conclusion(s): Bayesian estimation adds an essential measure of uncertainty to point estimates.

Trizano-Hermosilla et al. (2021) Estimator(s): $\alpha$, $\omega_{\text{hierarchical}}$, $\omega_{\text{total}}$, $\omega_{\text{limit}}$, GLB (P. H. Jackson & Agunwamba, 1977), GLB$_{\text{alg}}$ (Moltner & Revelle, 2016) • Data generation: Data generation procedure described, FA model, no software reported, no data validity check, no code provided • Replications: 500, number not justified • Conditions: Loading size for the general factor (0.4, 0.45, 0.5, 0.55, …, 0.8), loading size for specific factors (0.35, 0.45, 0.55), sample size (100, 150, 200, 250, 500, 1000), number of items (12, 24, 48), selection of conditions justified—to consider a range of conditions that can be observed in real situations • Benchmark: Population reliability specified for each coefficient, attributable both to general factor (general reliability) and to general and specific factors (total reliability) • Performance evaluation: M bias, SD of bias, minimum bias, maximum bias, coefficient of determination between population reliability and each coefficient, visualization used • Conclusion(s): $\alpha$, $\omega_{\text{total}}$, GLB (P. H. Jackson & Agunwamba, 1977), and GLB$_{\text{alg}}$ (Moltner & Revelle, 2016) are the least biased as estimators of total reliability; $\omega_{\text{hierarchical}}$ and $\omega_{\text{limit}}$ are the least biased estimates of general reliability.

References


Studies included after the additional search


**Excluded studies**

*Does not pertain to psychometric reliability*


*Text not available in English*


Monte Carlo simulation studies of reliability in psychometrics

Investigates rater agreement

Investigates reliability of profiles

Theoretical article

Does not investigate reliability estimators


Appendix C. Footnotes

Notes to Table 2

1. If multiple simulations were conducted that investigated an equal number of estimators per simulation, they were treated as a single unit for counting the number of investigated estimators. Only two studies investigated different number of estimators per simulation (one and three; one, one, and six) and they were left out for this display, which means there are 83 studies overall used for the display of frequencies and calculation of percentages for a certain number of investigated estimators.

2. It should be noted that the term factor represents a factor in the factorial design, not the number of dimensions that is potentially included as a factor in the design. Also, zero factors represent the cases in which a single empirical dataset was used for the data generation or certain specific conditions were generated without factor manipulation. If a study included multiple simulations, each of the simulations was treated as a separate unit for counting number of factors, which means there are 111 units overall used for the display of frequencies and calculation of percentages for a certain number of factors.

3. Although SD is descriptive, in the case of simulation studies it represents a performance measure, so if a study used SD of estimator values for performance evaluation, it was coded as a performance measure and not descriptive statistics.

Notes to Table 4

1. This does not contain obscure forms of $\alpha$, despite they are related to the formula presented by Guttman (1945). Obscured forms are treated separately since their inclusion in this category would confound the results.

2. Despite being a special case of $\alpha$ for dichotomous scales, it was displayed as a separate coefficient.

3. $\omega$ family represents any coefficient in the family being investigated in a particular study and contains $\omega_{\text{hierarchical}}$ and $\omega_{\text{categorical}}$, which are also displayed separately.

4. Despite FA is a special case of SEM, SEM-based coefficients are here discerned from the $\omega$ family due to subtle differences. The use of SEM-based coefficients requires additional modeling compared to the $\omega$ family in heterogeneous populations (e.g., Raykov & Marcoullides, 2015) and/or when unidimensionality does not hold, in which cases the researcher has to decide which model is the correct one (Savalei & Reise, 2019a) and possibly allow errors to correlate (Green & Yang, 2009). This is especially the case in longitudinal designs, for which an entire family of SEM-based coefficients was suggested (e.g., Laenen et al., 2009). Some researchers refer to all FA-based and SEM-based coefficients as $\omega$ (e.g., E. Cho, 2016).

5. Bias represents the mean (occasionally median) difference between the observed value and the benchmark value across replications. Absolute bias represents bias irrespective of its direction, whereas relative bias represents difference between an estimator and different values of benchmark that are comparable due to being on the equal scale (mostly percent). Precision represents the SD of estimator values across replications. Coverage rate represents the probability that an estimator value falls into confidence interval around the benchmark value. MSE represents the squared bias,
while root means square error represents square root of MSE. It is recommended to see papers cited in the introduction for details on these performance measures.

6. In some studies multiple validity check methods were used.

7. The first part of this section displays ranges of observed values, while the second part displays typical individual values (that appear at least five times). If a study reported using multiple numbers of replications, each was treated as a separate unit for counting number of replications, which means there are 87 units overall used for the display of frequencies and calculation of percentages for a certain number of replications.

8. Although item discrimination is basically equivalent to factor loading, it was coded as a separate factor since studies that included item discrimination as a factor were not focused on the measurement model type.