Statistical Properties of Lower Bounds and Factor Analysis Methods for Reliability Estimation

Julius M. Pfadt and Klaas Sijtsma

Abstract In this study, we compared the numerical performance of reliability coefficients based on classical test theory and factor analysis. We investigated the coefficients' divergence from reliability and their population values using unidimensional and multidimensional data generated from both an item response theory and a factor model. In addition, we studied reliability coefficients' performance when the tested model was misspecified. For unidimensionality, coefficients α, λ_2, and coefficient ω_u approximated reliability well and were almost unbiased regardless of the data-generating model. For multidimensionality, coefficient ω_t performed best with both data generating models. When the tested model was unidimensional but the data multidimensional, all coefficients underestimated reliability. When the tested model incorrectly assumed a common factor in addition to group factors but the data was purely multidimensional, coefficients ω_h and ω_t identified the underlying data structure well. In practice, we recommend researchers use reliability coefficients that are based on factor analysis when data are multidimensional; when data are unidimensional both classical test theory methods and factor analysis methods get the job done.

Keywords Classical test theory · Coefficient alpha · Coefficient lambda2 · Coefficient lambda4 · Coefficient omegaH · Coefficient omegaT · Coefficient omegaU · Factor analysis · Greatest lower bound

J. M. Pfadt (✉)
Ulm University, Department of Psychological Research Methods, Ulm University, Ulm, Germany
e-mail: julius.pfadt@uni-ulm.de

K. Sijtsma
Department of Methodology and Statistics TSB, Tilburg University, Tilburg, The Netherlands
e-mail: k.sijtsma@tilburguniversity.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 393, https://doi.org/10.1007/978-3-031-04572-1_5

1 Introduction

Most methods for reliability estimation fall into methods based on classical test theory (CTT; Lord & Novick, 1968) and factor analysis (FA; e.g., Jöreskog, 1971). We studied sampling properties of lower bound coefficients α, λ_2, λ_4, and the greatest lower bound (GLB) from CTT, and coefficients ω_u, ω_h, and ω_t from FA. We ran a simulation study assessing statistical properties of these estimators, using item response theory (IRT) and FA models to generate unidimensional and multidimensional data while controlling the reliability. As far as we know, this comparison is novel. We investigated two special cases in which the population model is a multidimensional FA-model, but the tested model is misspecified. Before we discuss the process of data generation, we describe the reliability coefficients α, λ_2, λ_4, *GLB*, ω_u, ω_h, and ω_t.

1.1 CTT Coefficients

Based on the CTT layout of the test score, $X = T + E$ (X: test score, sum of item scores; T: true score; E: random measurement error; see Sijtsma & Pfadt, 2021), we studied coefficients α, λ_2, λ_4, and the *GLB*. These coefficients approximate reliability, ρ, defined as the proportion of test-score variance that is true-score variance in a population; that is, $\rho = \sigma_T^2/\sigma_X^2$. The approximations are theoretical lower bounds to the reliability. The coefficients' equations are the following (for further information, see, e.g., Sijtsma & Van der Ark, 2021). For coefficient α, let σ_X^2 be the test score variance and let σ_j^2 be the variance of item j ($j = 1, \ldots, J$); then

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J}\sigma_j^2}{\sigma_X^2}\right). \tag{1}$$

For coefficient λ_2, let σ_{jk}^2 be the covariance between items j and k, then

$$\lambda_2 = 1 - \frac{\sum_{j=1}^{J}\sigma_j^2 - \sqrt{\frac{J}{J-1}\sum\sum_{j\neq k}\sigma_{jk}^2}}{\sigma_X^2}. \tag{2}$$

For coefficient λ_4, split a test in two item subsets without overlap and not necessarily equally sized, call this split partition P, and let σ_A^2 and σ_B^2 be the variances of the test scores on each subset. Then, coefficient α for this partition, $\alpha(P)$, equals: $\alpha(P) = 2 \cdot \left(1 - \frac{\sigma_A^2 + \sigma_B^2}{\sigma_X^2}\right)$. Coefficient λ_4 is the greatest value of $\alpha(P)$ across all partitions P; that is,

$$\lambda_4 = \max_P\left[\alpha(P)\right]. \tag{3}$$

Let $\mathbf{\Sigma}_X$ be the covariance matrix that is split into $\mathbf{\Sigma}_X = \mathbf{\Sigma}_T + \mathbf{\Sigma}_E$, where $\mathbf{\Sigma}_T$ contains the true score variances and diagonal $\mathbf{\Sigma}_E$ contains the error score variances. The estimates for $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_E$ are found in an iterative procedure where the trace of $\mathbf{\Sigma}_E$ is maximized while the matrices $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_E$ stay positive semi-definite (Woodhouse & Jackson, 1977). Then

$$GLB = 1 - \frac{tr\left(\mathbf{\Sigma}_E\right)}{\sigma_X^2}. \tag{4}$$

1.2 FA Coefficients

The FA approach to reliability is based on the assumption that the CTT model, $X = T + E$, can be substituted with the FA-model. Note that the models are different and thus define different reliability conceptions except for extreme cases. For items, the CTT model is $X_j = T_j + E_j$, with $T_j = \mathcal{E}\left(X_{jr}\right)$, r indexes independent replications of item j, comparable with the stochastic subject formulation of response probability in IRT (Holland, 1990). FA models item scores as $X_j = b_j + \sum_{q=1}^{Q} a_{jq}\xi_q + \delta_j$, with intercept b_j, latent variable ξ_q indexed q ($q = 1, \ldots, Q$), a_{jq} the loading of item j on latent variable ξ_q, and δ_j the residual consisting of unexplained item components and random measurement error. One may notice that different items have one or more latent variables in common, whereas in CTT, true scores unique to the item lack additional modeling. FA extracts one or more factors from the data. These factors predict the outcome variables, here, the items. The level of prediction is represented by loadings that link each item with one or more factors.

The FA approach to reliability is based on the assumption that the sum of the squared factor loadings approximates the true-score variance of items. The residual variances, the part of items the factor(s) cannot predict, substitute the error-score variance. Reliability is defined as the proportion of the test-score variance that is due to one or more common factors. For instance, the single-factor model (Spearman, 1904) describes the data matrix $\mathbf{X}$ of multivariate observations as

$$\mathbf{X} = \mathbf{g}\mathbf{c}^T + \mathbf{E}, \tag{5}$$

where $\mathbf{c}$ denotes the factor loadings on one common factor $\mathbf{g}$ (replacing general notation ξ) and $\mathbf{E}$ the matrix of residuals, the part of the item scores that the common factor cannot explain. Since the residuals are assumed independent, the covariance matrix of $\mathbf{E}$ is diagonal and has elements e_j representing residual variances**.** Coefficient ω_u (u for unidimensional; McDonald, 1999) equals

$$\omega_u = \frac{\left(\sum c\right)^2}{\left(\sum c\right)^2 + \sum e}. \tag{6}$$

When data are multidimensional, one can either estimate reliability for each subscale using ω_u or one can employ coefficients ω_h (h for hierarchical) and ω_t (t for total; Zinbarg et al., 2005). Therefore, consider the following multidimensional bi-factor model,

$$\mathbf{X} = \mathbf{g}\mathbf{c}^T + \mathbf{F}\mathbf{A}^T + \mathbf{E}, \tag{7}$$

where $\mathbf{A}$ denotes the $J \times Q$ loading matrix for the Q group factors collected in $\mathbf{F}$. The Q group factors are common to some items but not all. The residual variances e_j of the residual matrix $\mathbf{E}$ represent the part of the items that the common factor and the group factors cannot explain. Coefficient ω_h equals

$$\omega_h = \frac{\mathbf{c}^T\mathbf{c}}{\mathbf{c}^T\mathbf{c} + \mathbf{1}_Q^T\mathbf{A}^T\mathbf{A}\mathbf{1}_Q + \sum \boldsymbol{e}}, \tag{8}$$

where $\mathbf{1}_Q$ is a $Q \times 1$ sum vector. The coefficient describes the common factor saturation of a test in the presence of group factors. The value of coefficient ω_h addresses the question "how well does a multidimensional scale represent a common attribute". Coefficient ω_h is not an estimate of the reliability as indicated by coefficient ω_t, because coefficient ω_h does not relate all true score variance to the total variance of a test, but only the variance due to a general factor. Coefficient ω_t equals

$$\omega_t = \frac{\mathbf{c}^T\mathbf{c} + \mathbf{1}_Q^T\mathbf{A}^T\mathbf{A}\mathbf{1}_Q}{\mathbf{c}^T\mathbf{c} + \mathbf{1}_Q^T\mathbf{A}^T\mathbf{A}\mathbf{1}_Q + \sum \boldsymbol{e}}, \tag{9}$$

and describes the proportion of variance in the test that is due to a common attribute and specific attributes that pertain to subsets of items, which is the true-score variance. The loadings and residual variances for the ω-coefficients can be obtained from both a confirmatory factor model and an exploratory factor model.

2 Simulation Study

We compared reliability coefficients estimated in samples of simulated data with the reliability of the population model that generated the data. Oosterwijk et al. (2017) compared several lower bounds with population reliability by generating data from a two-dimensional graded response model (GRM; Samejima, 1968). From the GRM parameters, they computed the item true scores and then the reliability. Zinbarg et al. (2006) used a factor model to generate data for the evaluation of estimation methods for coefficient ω_h. Assuming the factor variance represents the true score variance, one can obtain the population reliability from the factor model parameters.

To rule out the possibility that the data generation process confounds the outcomes, we generated data both with the GRM and a factor model, and evaluated the lower bound coefficients α, λ_2, λ_4, and the *GLB*, and the FA-coefficients ω_u, ω_h, and ω_t. Furthermore, we investigated the ramifications of a misspecified model when estimating reliability coefficients. We generated data based on researchers (1) overlooking a scale's multidimensionality, and (2) incorrectly assuming presence of a common attribute.

2.1 Method

Data Generation from GRM

The data generation based on the GRM in this study is similar to the data generation in Oosterwijk et al. (2017). The GRM defines for each polytomous item j a slope parameter a_j and a location parameter b_{jx} for each item score x. The cumulative response probability of scoring at least x ($x = 0, \ldots, m$) on item j as a function of latent variable(s) θ_q ($q = 1, \ldots, Q$) collected in $\boldsymbol{\theta}$ is expressed as

$$P\left(X_j \geq x \mid \boldsymbol{\theta}\right) = \frac{\exp\left[\sum_{q=1}^{Q} a_{jq}\left(\theta_q - b_{jx}\right)\right]}{1 + \exp\left[\sum_{q=1}^{Q} a_{jq}\left(\theta_q - b_{jx}\right)\right]}. \tag{10}$$

The response probability of scoring exactly x on item j is given by

$$P\left(X_j = x \mid \boldsymbol{\theta}\right) = P\left(X_j \geq x \mid \boldsymbol{\theta}\right) - P\left(X_j \geq x + 1 \mid \boldsymbol{\theta}\right), \tag{11}$$

with $P(X_j \geq 0 | \boldsymbol{\theta}) = 1$ and $P(X_j > m | \boldsymbol{\theta}) = 0$ for response categories $0, \ldots, 4$. For our study, we chose $Q = 1$ and $Q = 3$. The model definition and data generation for the GRM largely follows Oosterwijk et al. (2017) but for clarity we reiterate most of it here.

Unidimensional Model To model a wide range of person parameters, we defined $\boldsymbol{\theta} = -5, -4.95, -4.9, \ldots, 4.9, 4.95, 5$, that is, 201 evenly spaced values in total that followed a standard normal distribution. Numbers of items were $J = 9, 18$, item scores were $x = 0, \ldots, 4$, and slope parameters $a_j \in U(1, 1.5)$. Item location parameters were $b_{jx} = \tau_j + \kappa_x$, with $\tau_j = (j-1)/(J-1)\ -.5$ and $\kappa_x = (-1.5, -0.5, 0.5, 1.5)^T$, with $x = 0, \ldots, 4$ for five item scores.

Multidimensional Model We chose $Q = 3$, identical equally-spaced person parameter vectors θ_1, θ_2, θ_3 as in the unidimensional case, and a multivariate normal distribution with means 0, variances 1, and correlations .3. Numbers of items and item scores were the same as for the unidimensional model. Multidimensionality was achieved by assigning slopes $a_{jq} \in U(1, 2)$ for latent variable θ_q to a third of the items and $a_{jq} = 0$ to the other items, and so on. For example, for $J = 9$, $\mathbf{a}_1 = (1.75, 1.25, 1.55, \mathbf{0})^T$, $\mathbf{a}_2 = (\mathbf{0}, 1.60, 1.04, 1.30, \mathbf{0})^T$, and $\mathbf{a}_3 = (\mathbf{0}, 1.67, 1.40, 1.28)$. Herein, vector $\mathbf{0}$ denotes a vector of zeros; for $\mathbf{a}_1$, $\mathbf{0}$ has six elements, for $\mathbf{a}_2$, both **0s** have three elements each, and for $\mathbf{a}_3$, $\mathbf{0}$ has six elements. Location parameters were defined as in the unidimensional model.

Population Reliability Reliability equals $\rho = (\mathbf{1}^T\mathbf{\Sigma}_T\mathbf{1})/(\mathbf{1}^T\mathbf{\Sigma}_X\mathbf{1})$. Matrix $\mathbf{\Sigma}_T$ differs from matrix $\mathbf{\Sigma}_X$ by the diagonal only, which for $\mathbf{\Sigma}_T$ contains the item true-score variances. Population values of coefficients α, λ_2, λ_4, the *GLB*, ω_{u}, ω_h and ω_t (for the multidimensional model) were computed from covariance matrix $\mathbf{\Sigma}_X$. $\mathbf{\Sigma}_X$ has diagonal item variances σ_j^2 and off-diagonal covariances σ_{jk}^2. Following Oosterwijk et al. (2017), we compute $\sigma_j^2 = \mathcal{E}\left(X_j^2\right) - \left[\mathcal{E}(X_j)\right]^2$, and $\sigma_{jk} = \mathcal{E}\left(X_jX_k\right) - \mathcal{E}\left(X_j\right)\mathcal{E}(X_k)$. Furthermore, $\mathcal{E}\left(X_j\right) = \sum_x xP\left(X_j = x\right)$, $\mathcal{E}\left(X_j^2\right) = \sum_x x^2P\left(X_j = x\right)$ and $\mathcal{E}\left(X_jX_k\right) = \sum_x\sum_y xyP\left(X_j = x, X_k = y\right)$, with $x = 0, \ldots, 4$.

Marginal probability $P(X_j = x)$ equals

$$P\left(X_j = x\right) = \sum_{\theta_q} P\left(\theta_q\right) P\left(X_j = x|\theta_q\right), \tag{12}$$

and joint probabilities $P(X_j = x, X_k = y)$ equal

$$P\left(X_j = x, X_k = y\right) = \sum_{\theta_q} P\left(\theta_q\right) P\left(X_j = x|\theta_q\right) P\left(X_k = y|\theta_q\right). \tag{13}$$

To obtain $\mathbf{\Sigma}_T$, we substitute the diagonal of $\mathbf{\Sigma}_X$ with the true item variances $\sigma_{T_j}^2$. We compute $\sigma_{T_j}^2$ from

$$\sigma_{T_j}^2 = \sum_{\theta_q} P\left(\theta_q\right)\left[\left(T_j|\theta_q\right) - \mathcal{E}\left(T_j\right)\right]^2, \tag{14}$$

with true scores $T_j \mid \theta_q = \sum_x P\left(X_j \geq x|\theta_q\right)$, and $\mathcal{E}\left(T_j\right) = \mathcal{E}\left(X_j\right)$. Probability $P(\theta_q)$ is computed as follows. First, for $Q = 1$, we compute the value of the probability density function of the standard normal distribution at each θ-value and then transform the resulting values to the zero-to-one probability scale by dividing each value by the sum of all values. Second, for $Q = 3$, the reference probability density function is the multivariate normal. Subsequently, the value of the density function is computed for each possible permutation of the three θ-vectors.

Data Generation First, for $Q = 1$, we drew N θ-values from a standard normal distribution and computed the cumulative and exact response probabilities from Eqs. (10) and (11) for each of the N θ-values. For $Q = 3$, we drew N triplets of θ-values from a multivariate normal with a specified correlation matrix ($\rho = .3$) and computed the required probabilities. Second, using the exact response probabilities, we randomly drew N scores on sets of J Items from a multinomial distribution. We scaled these scores ordinally, consistent with a five-point Likert-scale.

Data Generation from FA-Model

We created a covariance matrix implied by a particular factor model. The procedure differed between one- and three-factor models. We used the matrix as the data-generating covariance matrix.

Unidimensional model The covariance matrix the model implies is defined as $\mathbf{\Sigma}_{\mathrm{U}} = \mathbf{c}\phi\boldsymbol{c}^{\boldsymbol{T}} + \boldsymbol{\Psi}$, with ϕ as the variance of factor $\mathbf{g}$, and $\boldsymbol{\Psi}$ as the diagonal covariance matrix of $\mathbf{E}$ (Eq. 5; Bollen, 1989). We sampled standardized model loadings in $\mathbf{c}$ from $U(0.3, 0.7)$. Then, the residual variances in the diagonal matrix $\boldsymbol{\Psi}$ become $1 - c^2$. We assumed factor variance is $\phi = 1$. We computed parameters for coefficients α, λ_2, the *GLB*, and ω_u from the model-implied covariance matrix $\mathbf{\Sigma}_{U}$.

Multidimensional Model The multi-factor model was a second-order model for which we assumed three group factors each explaining a unique third of the items, and one general factor explaining the group factors. Group factor loadings came from a uniform distribution $U(.4, 1)$ and loadings of group factors on the general factor came from $U(.5, 1)$. Loadings were standardized, so that squared loadings together with the residual variances added to 1. General-factor variance equaled 1. Using the Schmid-Leiman transformation (Schmid & Leiman, 1957) we transformed group and general factor loadings to loadings of a bi-factor model (Eq. 7). Residual variances were the same. The model-implied covariance matrix was $\mathbf{\Sigma}_{\mathrm{M}} = \mathbf{\Lambda\Phi\Lambda}^{T} + \boldsymbol{\Psi}$ (Bollen, 1989). Matrix $\mathbf{\Lambda}$ contains the loadings of the items on both the group factors and the general factor, $\mathbf{\Phi}$ is the diagonal factor covariance matrix with *Q+1* entries that equal 1, and $\boldsymbol{\Psi}$ is a diagonal matrix containing the residual variances. We computed the parameters for α, λ_2, λ_4,the *GLB*, ω_h, and ω_t from the model implied covariance matrix $\mathbf{\Sigma}_{\mathrm{M}}$.

Population Reliability Reliability is defined as $\rho = \sigma_T^2/\sigma_X^2$. Assuming that squared factor loadings represent item true-score variances, population reliability equaled ω_u (see Eq. 6) for the unidimensional models and ω_t (see Eq. 9) for the multidimensional models.

Data Generation We drew random samples from a multivariate normal distribution with means of 0 and model-implied covariance matrices $\mathbf{\Sigma}_U$ and $\mathbf{\Sigma}_M$, respectively. The resulting data were then continuous.

Factor Analysis Method We estimated all factor models by means of a confirmatory factor analysis (CFA). To obtain coefficients ω_h and ω_t, we first performed a CFA with a second-order factor model and transformed the resulting loadings into bi-factor loadings, $\mathbf{c}$ and $\mathbf{A}$ (Eq. 7), using the Schmid-Leiman transformation.

In the simulation study, we used λ_4 as the population coefficient, and $\lambda_{4(.05)}$ as the sample estimate in the simulation runs, where $\lambda_{4(.05)}$ is the .05 quantile of a distribution of approximations to λ_4 that avoid having to consider all possible item splits, thus running into combinatorial problems (Hunt & Bentler, 2015). Coefficient $\lambda_{4(.05)}$ counteracts chance capitalization that sometimes leads to gross overestimation of population reliability.

Conditions

Numbers of items were $J = 9$, 18. Sample sizes were $N = 500$, 2000. Together with two data generation methods and two dimensionality conditions, 16 conditions resulted that were replicated 1000 times. In addition, we identified two misspecified models when estimating FA-coefficients ω_{u}, ω_h, and ω_t. First, we considered incorrectly assuming one factor, thus estimating coefficient ω_{u}, when the truth is a multi-factor model. We generated data using a second-order factor model with three group factors and a general factor, for $J = 12$ and $N = 1000$, with 1000 replications, and computed coefficient ω_{u}. Second, we considered incorrectly assuming multiple factors with an underlying common factor, thus computing coefficients ω_h and ω_t, when the true model is purely multi-factor and does not contain a common factor. The factor model had three orthogonal group factors. We assumed that coefficient ω_t equaled the population reliability and coefficient ω_h equaled zero as loadings on the general factor were zero. Number of items was $J = 12$, sample size was $N = 1000$, and number of replications was 1000. We estimated α, λ_2, λ_4, the *GLB*, ω_h, and ω_t.

Outcome Variables

We determined discrepancy, and the mean and standard error of bias. Discrepancy is the difference between parameters for reliability methods and reliability, for example, $\alpha - \rho$. Bias is the difference between the mean sample coefficient and its parameter value, for example, $\mathcal{E}\left(\hat{\alpha}\right) - \alpha$. Standard error is the standard deviation of estimates relative to the parameter, for example, $\sigma_{\hat{\alpha}} = \left(\mathcal{E}\left[\left(\hat{\alpha} - \alpha\right)^2\right]\right)^{\frac{1}{2}}$. We tested significance of the bias being different from zero.

2.2 Results

Unidimensional Models

Table 1 shows discrepancy, bias, standard error and significance results. Coefficients α, λ_2, λ_4, *GLB*, and ω_u showed similar results in both data generating scenarios. The discrepancy of λ_2, the *GLB*, and ω_u was small in all unidimensional conditions.

Table 1 Discrepancy, bias, and standard error (between parentheses) of several reliability methods for unidimensional models

	$J = 9$			$J = 18$		
		N			N	
		500	2000		500	2000
Coefficient	Discrepancy	Bias		Discrepancy	Bias	
	IRT-data					
	$\rho = .798$			$\rho = .889$		
α	−0.48	−0.66 (0.42)	−0.32 (0.21)	−0.21	−0.34 (0.23)	−0.06 (0.12)
λ_2	−0.29	00.06 (0.41)	−0.15 (0.21)	−0.12	−0.08 (0.23)	0.05 (0.11)
λ_4	−4.46	−8.31 (0.42)*	−7.39 (0.21)*	0	−8.05 (0.22)*	−5.95 (0.12)*
GLB	0	28.51 (0.38)*	14.12 (0.21)*	0	−27.10 (0.19)*	13.91 (0.11)*
ω_u	−0.23	−0.21 (0.41)	−0.21 (0.21)	−0.11	−0.13 (0.23)	−0.01 (0.12)
	FA-data					
	$\rho = .748$			$\rho = .859$		
α	−4.61	−1.17 (0.53)*	−0.57 (0.27)*	−2.16	−0.46 (0.30)	−0.07 (0.15)
λ_2	−0.87	0.04 (0.52)	0.31 (0.26)*	−0.22	−0.20 (0.29)	−0.09 (0.15)
λ_4	0	−15.88 (0.53)*	−11.78 (0.28)*	0	−3.12 (0.28)*	−4.75 (0.15)*
GLB	0	−35.11 (0.50)*	−17.66 (0.26)*	0	33.86 (0.24)*	17.42 (0.13)*
ω_u	0	−0.47 (0.53)	−0.42 (0.26)	0	−0.13 (0.29)	0.00 (0.15)

Note. Significance is indicated with *. Table entries are transformed and rounded for better interpretation; real values are obtained by multiplying entries by 10^{-3}, e.g., the discrepancy for α ($J = 9$; IRT-data) is $-0.48 \times 0.001 = -0.00048$. Discrepancy for λ_4 was $\lambda_4 - \rho$, bias was estimated using $\lambda_{4(.05)}$

The discrepancy of coefficient λ_4 improved considerably with a larger number of items. Discrepancy was negative for all coefficients, a desirable result. Mean bias of coefficients α, λ_2, and ω_u was relatively small. Table 1 shows that the discrepancy of the *GLB* is almost equal to 0, but its bias is largely positive, a finding consistent with results reported by Oosterwijk et al. (2017). Estimate $\lambda_{4(.05)}$ underestimated population value, λ_4. Increase in sample size resulted in better performance for all coefficients. Except for the *GLB*, an increase in the number of items led to smaller bias. Except for λ_4 and the *GLB*, the coefficients' performance was satisfactory across all unidimensional conditions.

Multidimensional Models

Table 2 shows that all coefficients had smaller discrepancy and bias as samples grew larger and, except for the *GLB*, results improved as the number of items grew. Discrepancy was highly similar for both data generation procedures. As expected, discrepancy of lower bounds α and λ_2 was much larger for the multidimensional data than for the unidimensional data. Coefficient λ_4 showed an unexpectedly large discrepancy with the multidimensional IRT-data and considerable negative bias throughout all multidimensional conditions. The *GLB* had very small discrepancy but an expectedly large bias. As expected, an increase in the number of items

Table 2 Discrepancy, bias, and standard error (between parentheses) of several reliability methods for multidimensional models

	$J = 9$			$J = 18$		
		N			N	
		500	2000		500	2000
Coefficient	Discrepancy	Bias		Discrepancy	Bias	
	IRT-data					
	$\rho = .738$			$\rho = .853$		
α	−82.13	−0.38 (0.73)	−0.18 (0.36)	−44.61	−0.77 (0.39)*	−0.39 (0.20)
λ_2	−67.82	−1.41 (0.65)*	−0.24 (0.32)	−36.14	−0.23 (0.35)	−0.12 (0.18)
λ_4	−31.74	−16.32 (0.59)*	−14.94 (0.31)*	−0.04	−13.72 (0.31)*	−8.65 (0.16)*
GLB	−0.07	26.15 (0.52)*	−12.08 (0.28)*	0	32.34 (0.25)*	16.37 (0.15)*
ω_h	−325.91	7.01 (1.39)*	1.82 (0.69)*	−376.23	3.41 (1.23)*	−0.18 (0.64)
ω_t	−0.10	0.70 (0.55)	0.05 (0.27)	−0.04	−0.37 (0.29)	−0.24 (0.15)
	FA-data					
	$\rho = .872$			$\rho = .917$		
α	−59.19	−1.46 (0.40)*	−0.59 (0.21)*	−29.84	−0.36 (0.23)	−0.20 (0.12)
λ_2	−49.93	−0.67 (0.36)	−0.37 (0.19)*	−24.41	−0.10 (0.22)	−0.06 (0.11)
λ_4	−13.18	−30.18 (0.34)*	−26.8 (0.17)*	−0.28	−11.68 (0.17)*	−11.99 (0.08)*
GLB	0	11.74 (0.26)*	−5.65 (0.14)*	0	−18.14 (0.15)*	9.17 (0.08)*
ω_h	−217.61	−0.98 (0.83)	−0.87 (0.42)*	−212.22	−0.42 (0.69)	−0.20 (0.35)
ω_t	0	−0.55 (0.27)*	−0.24 (0.14)	0	−0.09 (0.17)	−0.04 (0.08)

Note. Significance is indicated with *. Table entries are transformed and rounded for better interpretation; real values are obtained by multiplying entries by 10^{-3}, e.g., the discrepancy for α ($J = 9$; IRT-data) is $-82.1 \times 0.001 = -0.0821$. Discrepancy for λ_4 was $\lambda_4 - \rho$, bias was estimated using $\lambda_{4(.05)}$

produced a larger bias for the *GLB*, because capitalization on chance increases with the number of items.

Results for coefficient ω_h were not consistent with results for the other estimators. The population value of the coefficient ω_h was much lower than the population reliability. This was expected, given that ω_h indicates how well a common attribute is represented irrespective of the real factor structure. The difference between coefficients ω_h and ω_t indicates the presence of multidimensionality. Coefficient ω_h performed well with the FA-data, but with the IRT-data bias was positive, meaning it overestimated the population ω_h.

Coefficient ω_t performed well across all multidimensional conditions. It had negligible discrepancy (by definition, zero with the FA-data) and small mean bias across all conditions.

Misspecified Models

In the first case, the misspecification occurred by estimating the wrong coefficient, ω_u, which is suited for one-factor data expect data were in fact multi-factorial. The population value of coefficient ω_u was far from the population reliability (Table 3).

Table 3 Discrepancy, bias, and standard error (between parentheses) of several reliability methods for misspecified models

Coefficient	Discrepancy	Bias
	Case (1), $\rho = .899$	
ω_u	−42.18	−0.80 (0.22)*
	Case (2), $\rho = .807$	
α	−155.43	−1.23 (0.52)*
λ_2	−102.09	−0.15 (0.37)
λ_4	−2.95	−25.57 (0.32)*
GLB	0	19.88 (0.27)*
ω_h	−805.02	155.67 (3.76)*
ω_t	0	0.13 (0.27)

Note. Case (1): Multi-factor population model and data; computations assumed unidimensionality. Case (2): Multi-factor population model with group factor but no common factor, and data; computations assumed a common factor. Significance is indicated with a *. Table entries were transformed and rounded for better interpretation; real values are obtained by multiplying entries by 10^{-3}, e.g., discrepancy for ω_u is $-42.18 \times 0.001 = -0.04218$. $J = 12$ items and $N = 1000$

Subsequently, the estimates for ω_u were far off. In the second case, multi-factor data with a common factor was incorrectly assumed when the model generating the data contained only group factors but no common factor. The discrepancy of coefficients α and λ_2 was quite large, mirroring the multidimensionality of the data (Table 3). Coefficient λ_4 and the *GLB* had small discrepancy. The discrepancy of coefficient ω_h was huge, meaning the coefficient properly identified the absence of a common attribute. Because data were noisy, ω_h had considerable bias. Coefficient ω_t showed small bias. Its discrepancy was 0 since we used a factor model to generate the data. Arguably, a cautious researcher should always check model fit before estimating reliability coefficients that assume a certain structure of the data.

3 Discussion

The population values of the reliability methods were all fairly close to the population reliability with unidimensional data, which changed when data were multidimensional. Most coefficients had small bias in almost all conditions, except for the positive bias of the *GLB* and the negative bias of λ_4. All coefficients did well with unidimensional data. For multidimensional data, coefficients α and λ_2 on average underestimated reliability, while other methods were closer to reliability. For high reliability, coefficients α and λ_2 had high values, albeit somewhat smaller than the true reliability. The question is whether in this situation one should rather estimate reliability for each dimension separately. Among other things, this depends on the practical use of the test when it is sensible to distinguish different attributes or different aspects of the same attribute. In general, FA-coefficients performed very well. Coefficient λ_4 and the corrected estimate $\lambda_{4(.05)}$ were not satisfactory, because the discrepancy between λ_4 and population reliability was often larger than expected and the bias of $\lambda_{4(.05)}$ was too large to distinguish the coefficient from other lower bounds such as α.

The goal of the simulation study was investigating whether the reliability methods performed differently with discrete data generated by IRT and continuous data generated by FA. Does this difference in data types cause problems when comparing reliability methods results? We argue it does not, because different data types only present another hurdle the coefficients have to take rendering their performance evaluation more interesting. Regarding our simulation outcomes (discrepancy and bias), we found that the methods performed equally well with ordinal IRT data as with continuous FA data.

A limitation to our study was that we considered point estimation, but interval estimation, which is not common practice yet, may be more informative. Recent studies have shown that with unidimensional and multidimensional data, Bayesian credible intervals for coefficients α, λ_2, the GLB, ω_u, ω_h, and ω_t perform well (Pfadt, van den Bergh, & Moshagen, 2021a, b). We assume that the credible intervals of the reliability coefficients relate to population reliability in the same way as the population values of the coefficients do (as denoted by the discrepancy values we found).

In addition to the dominant CTT and FA reliability methods, less well-known methods based on IRT (Holland & Hoskens, 2003; Kim, 2012) and generalizability theory (GT; e.g., Brennan, 2001) exist. In IRT, use of typical IRT reliability methods is rare given the focus on the scale-dependent information function, which proves to be a powerful tool in IRT applications, such as adaptive testing and equating. GT provides an attempt to incorporate the influence of different facets of the test design and environment in the estimation of reliability. Suppose one studies the effect of test version and score rater on test performance. This requires a design with factors persons (i), test versions (t), and raters (r). Item scores are decomposed into person effect (ν_i), test effect (ν_t), and rater effect (ν_r), interaction effects, and a residual effect (Δ_{itr}), comparable with an ANOVA layout, so that $X_{itr} = \mu + \nu_i + \nu_t + \nu_r + \nu_{it} + \nu_{ir} + \nu_{tr} + \Delta_{itr}$. Reliability methods, called generalizability and dependability methods identify variance sources that affect relative and absolute person ordering, respectively, and correct for other, irrelevant sources. The GT approach provides a different perspective relevant to some research contexts where richer data are available and is worth pursuing in future research.

To conclude, when researchers have unidimensional data, the choice of a reliability coefficient is mostly arbitrary (if λ_4 and the *GLB* are discarded). With multidimensional data, the use of a factor model coefficient is encouraged, but lower bounds such as α prevent researchers from being too optimistic about reliability.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc.. https://doi.org/10.1002/9781118619179

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*(4), 577–601. https://doi.org/10.1007/BF02294609

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68*(1), 123–149. https://doi.org/10.1007/BF02296657

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

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

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika, 77*(1), 153–162. https://doi.org/10.1007/s11336-011-9238-0

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

McDonald, R. P. (1999). *Test theory: A unified treatment* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410601087

Oosterwijk, P. R., Van der Ark, L. A., & Sijtsma, K. (2017). Overestimation of reliability by Guttman's λ4, λ5, and λ6 and the greatest lower bound. In L. A. van der Ark, S. Culpepper, J. A. Douglas, W.-C. Wang, & M. Wiberg (Eds.), *Quantitative psychology research: The 81th annual meeting of the psychometric society 2016* (pp. 159–172). Springer. https://doi.org/10.1007/978-3-319-56294-0_15

Pfadt, J. M., van den Bergh, D., Moshagen, M. (2021a). *The reliability of multidimensional scales: A comparison of confidence intervals and a Bayesian alternative* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/d3gfs

Pfadt, J. M., van den Bergh, D., Sijtsma, K., Moshagen, M., & Wagenmakers, E. -J. (2021b). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2021.1891855

Samejima, F. (1968). *Estimation of latent ability using a response pattern of graded scores*. Educational Testing Service.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53–61. https://doi.org/10.1007/BF02289209

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

Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429112447

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

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42*(4), 579–591. https://doi.org/10.1007/bf02295980

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

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