Overestimation of reliability by Guttman’s $\lambda_4$, $\lambda_5$, and $\lambda_6$, and the greatest lower bound

Oosterwijk, P.R.; van der Ark, L.A.; Sijtsma, K.

Published in: Quantitative psychology

DOI: 10.1007/978-3-319-56294-0_15

Citation for published version (APA):
Overestimation of Reliability by Guttman’s $\lambda_4$, $\lambda_5$, and $\lambda_6$ and the Greatest Lower Bound

Pieter R. Oosterwijk, L. Andries van der Ark, and Klaas Sijtsma

Abstract For methods using statistical optimization to estimate lower bounds to test-score reliability, we investigated the degree to which they overestimate true reliability. Optimization methods do not only exploit real relationships between items but also tend to capitalize on sampling error and do this more strongly as sample size is smaller and tests are longer. The optimization methods were Guttman’s $\lambda_4$, $\lambda_5$, and $\lambda_6$ and the greatest lower bound to the reliability (GLB). Method $\lambda_2$ was used as benchmark. We used a simulation study to investigate the relation of the methods’ discrepancy, bias, and sampling error with the proportion of simulated data sets in which each method overestimated true test-score reliability. Method $\lambda_4$ and the GLB often overestimated test-score reliability. When sample size exceeded 250 observations, methods $\lambda_2$, $\lambda_5$, and $\lambda_6$ provided reasonable to good statistical results, in particular when data were two-dimensional. Benchmark method $\lambda_2$ produced the best results.

Keywords Chance capitalization of reliability optimization methods • Greatest lower bound to the reliability • Guttman’s $\lambda_2$ • Guttman’s $\lambda_4$ • Guttman’s $\lambda_5$ • Guttman’s $\lambda_6$ • Reliability overestimation
1 Introduction

Reliability quantifies the degree to which test scores can be repeated under identical test administration conditions, in which neither the examinee (with respect to the measured attribute) nor the test (with respect to content) has changed. Perfect repeatability is hampered by random influences beyond the test administrator’s control affecting test scores, causing test scores obtained in different administrations to be different. Reliability is the correlation between two test scores, denoted \( X \) and \( X' \), obtained independently in a group of examinees (Lord and Novick 1968, p. 46), and is denoted \( \rho_{XX'} \). Researchers usually collect item scores based on one test administration and use methods to approximate \( \rho_{XX'} \) based on this single data set. These methods usually produce lower bounds to \( \rho_{XX'} \).

Some of the approximation methods optimize a criterion based on the data in an effort to approximate \( \rho_{XX'} \) as close as possible. However, because they capitalize on sample characteristics, smaller samples cause methods to overestimate \( \rho_{XX'} \) more often and to a greater extent. Increasing the number of items also invites more chance capitalization. Overestimation is undesirable, because test users must be able to rely on the reported reliability estimate within the limits of statistical uncertainty, hence not providing values that are systematically too high.

Reliability overestimation has received little attention thus far. We study the degree to which four reliability methods using optimization overestimate \( \rho_{XX'} \). The methods are \( \lambda_4 \), \( \lambda_5 \), and \( \lambda_6 \) (Guttman 1945) and the greatest lower bound to the reliability (GLB; Bentler and Woodward 1980). We recommend which method to use for reliability estimation.

2 Test-Score Reliability

Of the methods using one data set to estimate reliability, coefficient \( \alpha \) (Cronbach 1951) is the most popular but not the best (Sijtsma 2009). Usually, these methods determine reliability based on the variance-covariance matrix of the items constituting the test. Many methods alternative to coefficient \( \alpha \) have been proposed (e.g., Bentler and Woodward 1980; Guttman 1945; Jackson and Agunwamba 1977; Kuder and Richardson 1937; Ten Berge and Zegers 1978). Sijtsma and Van der Ark (2015) also discuss methods based on factor analysis and generalizability theory. Guttman’s \( \lambda_4 \), \( \lambda_5 \), and \( \lambda_6 \) and the GLB employ a series of consecutive steps to optimize a method-dependent formal criterion defined on the sample data, resulting in an optimal value when the criterion is satisfied. Optimization methods are known to capitalize on chance, the more so when samples are smaller and tests are longer, hence producing more and greater overestimation effects. Such effects are unknown for Guttman’s \( \lambda_4 \), \( \lambda_5 \), and \( \lambda_6 \) and the GLB.
2.1 Classical Reliability Definition

Test score $X$ is the sum of $J$ item scores, denoted $X_j$, with $j = 1, \ldots, J$, such that $X = \sum_{j=1}^{J} X_j$. CTT assumes that test score $X$ can be decomposed in an unobservable, true-score part, denoted $T$, and an unobservable, random measurement error, denoted $E$, such that

$$X = T + E. \quad (1)$$

The score decomposition can be applied to any measurement value, for example, an individual item score; then, $X_j = T_j + E_j$, and Eq. (1) equals

$$\sum_{j=1}^{J} X_j = \sum_{j=1}^{J} T_j + \sum_{j=1}^{J} E_j. \quad (2)$$

Because measurement error $E$ is random, error correlates 0 with other variables, and it follows that (a) the group variance of the test score equals

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2, \quad (3)$$

and (b) error variance for the test score equals the sum of the item error variances; that is,

$$\sigma_E^2 = \sum_{j=1}^{J} \sigma_{E_j}^2. \quad (4)$$

Measurements $X$ and $X'$ are parallel if (1) for each examinee, $T_i = T'_i$; hence, at the group level, $\sigma_T^2 = \sigma_T'^2$, and (2) at the group level, $\sigma_X^2 = \sigma_X'^2$. Lord and Novick (1968, p. 61) showed that $\rho_{XX'}$ can be written as

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}. \quad (5)$$

Using Eq. (4), we can write the right-hand side of Eq. (5) as

$$\rho_{XX'} = 1 - \frac{\sum_{j=1}^{J} \sigma_{E_j}^2}{\sigma_X^2}. \quad (6)$$

Because, in practice, parallel measures usually are unavailable and because Eqs. (5) and (6) contain too many unknowns, $\rho_{XX'}$ is estimated using the data from one test administration. Methods $\lambda_4$, $\lambda_5$, and $\lambda_6$ and the GLB each seek a unique upper bound for the numerator in Eq. (6), $\sum_{j=1}^{J} \sigma_{E_j}^2$, and thus find a lower bound for $\rho_{XX'}$. 
We use the following notation. Let $\sigma_{jk}$ be the inter-item covariance. One can derive that $\sigma_{TjTk} = \sigma_{jk}$, and by definition $\sigma_{EjEk} = 0, j \neq k$. If Eq. (3) is rewritten for individual items, we have

$$\sigma^2_{Xj} = \sigma^2_{Tj} + \sigma^2_{Ej}.$$  \hspace{1cm} (7)

Covariance matrices $\Sigma_X$ and $\Sigma_T$ are symmetrical and have order $J \times J$, and $\Sigma_E$ is diagonal and has order $J \times J$, so that

$$\Sigma_X = \Sigma_T + \Sigma_E.$$ \hspace{1cm} (8)

Matrix $\Sigma_X$ is positive definite (pd); that is, for any vector $u$ of size $J$, we have $u^T \Sigma_X u > 0$ (i.e., $\Sigma_X$ has a positive determinant); $\Sigma_T$ and $\Sigma_E$ are positive semi-definite (psd), so that $u^T \Sigma_X u \geq 0$ (i.e., determinants are nonnegative).

The derivation of the methods $\lambda_4$, $\lambda_5$, and $\lambda_6$ and also benchmark $\lambda_2$ can be found with Guttman (1945) and Jackson and Agunwamba (1977), and for the GLB we refer the reader to Bentler and Woodward (1980). Because derivations are known, we only provide results.

### 2.2 Methods $\lambda_4$, $\lambda_5$, and $\lambda_6$, GLB, and Benchmark $\lambda_2$

#### 2.2.1 Method $\lambda_4$

Method $\lambda_4$ is based on splitting the $J$-item test in two parts, not necessarily of equal length, and finds the split that minimizes an appropriate upper bound for $\sum_{j=1}^{J} \sigma^2_{Ej}$ in Eq. (6) and consequently a lower bound for $\rho_{XX'}$. Here, we define the upper bound for $\sum_{j=1}^{J} \sigma^2_{Ej}$ typical of method $\lambda_4$. Let $u$ only have elements equal to either $+1$ or $-1$ so that $u$ selects items in either of the two test parts of a particular test split. It can be shown that

$$u^T \Sigma_X u = u^T \Sigma_T u + u^T \Sigma_E u.$$ \hspace{1cm} (9)

from which it follows that

$$u^T \Sigma_E u = \sum_{j=1}^{J} \sigma^2_{Ej} \leq u^T \Sigma_X u.$$ \hspace{1cm} (10)

The right-hand side of Eq. (10) provides an upper bound for $\sum_{j=1}^{J} \sigma^2_{Ej}$, and method $\lambda_4$ finds the vector $u$ that minimizes $u^T \Sigma_X u$, so that

$$\lambda_4 = \max_u \left( 1 - \frac{u^T \Sigma_X u}{\sigma^2_X} \right).$$ \hspace{1cm} (11)
Overestimation of Reliability by Guttman’s $\lambda_4$, $\lambda_5$, and $\lambda_6$ and the Greatest ...

Because $u$ and $-u$ provide the same value for $u^T \Sigma_X u$, and because vectors $u$ containing only $+1$s or only $-1$s do not refer to a test split, $2^{J-1} - 2$ vectors and corresponding products $u^T \Sigma_X u$ remain to find $\lambda_4$. For test length $J < 20$, one can try all vectors $u$ within reasonable computing time, and for test length $J \geq 20$, we refer to a procedure proposed by Benton (2015).

2.2.2 Method $\lambda_5$

From $\Sigma_T$ being psd, it follows that every principal submatrix of $\Sigma_T$ also has a nonnegative determinant, so that, for example, $\sigma_{Tj}^2 \sigma_{Tk}^2 \geq \sigma_{jk}^2$, all $j \neq k$. One can use this result to derive for a fixed column $k$ of $\Sigma_T$ containing $J - 1$ covariances $\sigma_{jk}$ ($k \neq j$) and ignoring $\sigma_{Xk}^2$ that (noticing that $\sum_{k \neq j}$ produces summation across index $k, k \neq j$)

$$\sum_{j=1}^{J} \sigma_{Tj}^2 \geq 2 \left( \sum_{k \neq j} \sigma_{jk}^2 \right)^{\frac{1}{2}}.$$ (12)

From Eq. (7) it follows that

$$\sum_{j=1}^{J} \sigma_{Xj}^2 = \sum_{j=1}^{J} \sigma_{Tj}^2 + \sum_{j=1}^{J} \sigma_{Ej}^2,$$ (13)

which implies

$$\sum_{j=1}^{J} \sigma_{Ej}^2 \leq \sum_{j=1}^{J} \sigma_{Xj}^2 - 2 \left( \sum_{k \neq j} \sigma_{jk}^2 \right)^{\frac{1}{2}}.$$ (14)

The right-hand side of Eq. (14) provides another upper bound for $\sum_{j=1}^{J} \sigma_{Ej}^2$, and one is free to choose the column $k$ that minimizes this upper bound, hence finds a lower bound for $\rho_{XX'}$. To find $\lambda_5$, let $k$ vary across each of the $J$ columns of $\Sigma_T$ and define

$$\lambda_5 = 1 - \frac{\sum_{j=1}^{J} \sigma_{Xj}^2 - \max_k \left( 2 \left( \sum_{k \neq j} \sigma_{jk}^2 \right)^{\frac{1}{2}} \right)}{\sigma_X^2}.$$ (15)

2.2.3 Method $\lambda_6$

Method $\lambda_6$ is based on the multiple regression of each of the $J$ item scores $X_j$ on the other $J - 1$ item scores. By minimizing the residual variance of the model, multiple
regression finds the regression weights for each of the \( J - 1 \) items. The residual variance of item \( j, \sigma^2_{\varepsilon_j} \), is an upper bound to the measurement error variance for item \( j, \sigma^2_{E_j} \); that is, \( \sigma^2_{E_j} \leq \sigma^2_{\varepsilon_j} \), and adding across the \( J \) items, we obtain

\[
\sum_{j=1}^{J} \sigma^2_{E_j} \leq \sum_{j=1}^{J} \sigma^2_{\varepsilon_j}
\]

(Jackson and Agunwamba 1977). Thus, the right-hand side of Eq. (16) provides yet another upper bound for the numerator in Eq. (6), which is \( \sum_{j=1}^{J} \sigma^2_{E_j} \). Replacing the numerator in Eq. (6) by the right-hand side of Eq. (16), produces a lower bound to the reliability,

\[
\lambda_6 = 1 - \frac{\sum_{j=1}^{J} \sigma^2_{\varepsilon_j}}{\sigma^2_X}.
\]

For estimation of \( \lambda_6 \) using covariance matrix \( \Sigma_X \), see Jackson and Agunwamba (1977) and Oosterwijk et al. (2016).

### 2.2.4 Greatest Lower Bound

Numerous pairs of different matrices \( \Sigma_T \) and \( \Sigma_E \) produce the same \( \Sigma_X \); see Eq. (8). Let \( tr(\Sigma_E) = \sum_{j=1}^{J} \sigma^2_{\varepsilon_j} \), and let \( \tilde{\Sigma}_E \) be the matrix of error variances for which the trace is maximized, provided that \( \tilde{\Sigma}_E \) is psd, and \( \tilde{\Sigma}_T \) is the corresponding covariance matrix of the item true scores, such that \( \Sigma_X = \tilde{\Sigma}_T + \tilde{\Sigma}_E \). Reliability [Eq. (5)] can be written as

\[
\rho_{XX'} = 1 - \frac{tr(\Sigma_E)}{\sigma^2_X},
\]

and the GLB is obtained by replacing \( tr(\Sigma_E) \) with \( tr(\tilde{\Sigma}_E) \), so that

\[
GLB = 1 - \frac{tr(\tilde{\Sigma}_E)}{\sigma^2_X}.
\]

The GLB algorithm used in this chapter is due to Bentler and Woodward (1980). If the \( J \) items or the test parts in which the test is divided are essential tau-equivalent (Lord and Novick 1968, p. 90), then \( GLB = \rho_{XX'} \). When essential tau-equivalence does not hold, the GLB provides the lowest possible reliability given the data, and \( GLB < \rho_{XX'} \), but other methods provide lower values, hence, smaller lower bounds, and the GLB thus is the greatest lower bound (Jackson and Agunwamba 1977).
2.2.5 Benchmark Method $\lambda_2$

Guttman (1945) derived three methods, denoted $\lambda_1$, $\lambda_2$, and $\lambda_3$ (equal to coefficient $\alpha$), which all use the inter-item covariances but do not optimize a statistical criterion and thus are not expected to capitalize on chance. Hence, in principle, each could serve as a benchmark for the methods $\lambda_4$, $\lambda_5$, and $\lambda_6$ and GLB. The relationship between the three methods and the reliability is

$$\lambda_1 < \lambda_3 (= \alpha) \leq \lambda_2 \leq \rho_{XX'}.$$  \hspace{1cm} (20)

In general, $\lambda_1$ is considered practically useless, and Sijtsma (2009) and Oosterwijk et al. (2016) have recommended using $\lambda_2$ rather than $\lambda_3$. Hence, we used $\lambda_2$ as benchmark for methods $\lambda_4$, $\lambda_5$, and $\lambda_6$ and GLB. Method $\lambda_2$ equals

$$\lambda_2 = 1 - \frac{\sum_{j=1}^{J} \sigma_{XX'}^2 - \left( \frac{\sum_{j=1}^{J} \sum_{j\neq k} \sigma_{jk}^2}{\sigma_{XX'}^2} \right)^{\frac{1}{2}}}{\sigma_{XX'}^2}.$$  \hspace{1cm} (21)

2.3 Knowledge About Reliability Methods in Samples

For $N < 1000$ and $J > 10$, the GLB is positively biased relative to $\rho_{XX'}$ (Shapiro and Ten Berge 2000; Ten Berge and Soćan 2004). Under particular conditions, method $\lambda_4$ has values similar to the GLB (Jackson and Agunwamba 1977; Ten Berge and Soćan 2004). Hence, method $\lambda_4$ has almost the same bias relative to $\rho_{XX'}$ as the GLB; Benton (2015) found that method $\lambda_4$ is biased when $N < 3000$. Samples this size are common in the social and the behavioral sciences, and results with respect to chance capitalization are needed for smaller samples, not only for method $\lambda_4$ and the GLB but also for $\lambda_5$ and $\lambda_6$. We used a simulation study to assess this problem.

3 Method

3.1 Population Model

Data were simulated using the two-dimensional graded response model (De Ayala 2009, pp. 275–305). The two-dimensional graded response model expresses the probability of scoring at least $x$ on item $j$ as a function of latent variable $\theta$, item location parameters $\beta_{jx}$, and item discrimination parameters $\alpha_j$, such that

$$P(X_j \geq x|\theta_1, \theta_2) = \frac{\exp[\alpha_{j1}(\theta_1 - \beta_{jx}) + \alpha_{j2}(\theta_2 - \beta_{jx})]}{1 + \exp[\alpha_{j1}(\theta_1 - \beta_{jx}) + \alpha_{j2}(\theta_2 - \beta_{jx})]}.$$  \hspace{1cm} (22)
Latent variables $\theta_1$ and $\theta_2$ had 101 equidistant values ($-5, -4.9, -4.8, \ldots, 5$) and were approximately bivariate normally distributed with mean 0, variance 1, and correlation $\rho_{\theta_1\theta_2}$. Joint probability is denoted $P(\theta_1, \theta_2)$. We assumed that the test consisted of $J$ items with five ordered item scores, $x = 0, \ldots, 4$.

Item location parameters $\beta_{jx}$ consisted of an item-specific part $\tau_j$ and a category-specific part $\kappa_x$, such that $\beta_{jx} = \tau_j + \kappa_x$. We chose $\tau_j = (j-1)/(J-1)-0.5$ and $\kappa_x = -0.75 + 0.5x$ and computed the item location parameters $\beta_{jx}$ for $j = 1, \ldots, 5; x = 1, \ldots, 4$. For five items, this resulted in $\tau = (-0.5, -0.25, 0, 0.25, 0.5)$, $\kappa = (-0.75, -0.25, 0.25, 0.75)$, and 20 $\beta$ values, which are readily computed.

Discrimination parameters $\alpha_j$ differed across latent variables. For $\theta_1$, $\alpha_{j1} = 1.6$ for the odd-numbered items and $\alpha_{j1} = 0$ for the even-numbered items. For $\theta_2$, $\alpha_{j2} = 1.6$ for the even-numbered items and $\alpha_{j2} = 0$ for the odd-numbered items. For 10 and 15 items, each next 5-tuple of items had the same discrimination parameters as the first 5-tuple; that is, using math operation modulo, for $k = j \mod 5$, for $j > 5$, one finds $\alpha_{j1} = \alpha_{k1}$ and $\alpha_{j2} = \alpha_{k2}$.

Equation (22) was used to compute covariance matrix $\Sigma_T$, so as to obtain the numerator of $\rho_{XX'}$ [Eq. (5)]. Because $\sigma_{T_j T_k} = \sigma_{j k}$, we only require the item true-score variances, $\sigma^2_{j j}$. It also may be noted that $E(T_j) = E(X_j)$. First, for fixed values of $\theta_1$ and $\theta_2$, the true score of item $j$ equals

$$T_j|\theta_1, \theta_2 = \sum_x P(X_j \geq x|\theta_1, \theta_2).$$

(23)

Second, the true-score variance of item $j$ then equals

$$\sigma^2_{T_j} = \sum_{\theta_1} \sum_{\theta_2} P(\theta_1, \theta_2) \left[ T_j|\theta_1, \theta_2 - E(T_j) \right]^2. \tag{24}$$

The conditional probability of obtaining item score $x$ on item $j$ equals

$$P(X_j = x|\theta_1, \theta_2) = P(X_j \geq x|\theta_1, \theta_2) - P(X_j \geq x+1|\theta_1, \theta_2).$$

(25)

Equation (25) was used to compute covariance matrix $\Sigma_X$, so as to obtain the denominator of $\rho_{XX'}$ [Eq. (5)], methods $\lambda_4$, $\lambda_5$, and $\lambda_6$, the GLB, and benchmark method $\lambda_2$. First, given discrete values for the latent variables, manifest marginal probabilities $P(X_j = x)$ and joint probabilities $P(X_j = x, X_k = y)$ were computed using

$$P(X_j = x) = \sum_{\theta_1} \sum_{\theta_2} P(\theta_1, \theta_2) P(X_j = x|\theta_1, \theta_2) \tag{26}$$

and

$$P(X_j = x, X_k = y) = \sum_{\theta_1} \sum_{\theta_2} P(\theta_1, \theta_2) P(X_j = x|\theta_1, \theta_2) P(X_k = y|\theta_1, \theta_2). \tag{27}$$
respectively. Second, the following expected values were computed using Eqs. (26) and (27): \( \mathcal{E}(X_j) = \sum_x xP(X_j = x) \), \( \mathcal{E}(X_j^2) = \sum_x \sum_y xyP(X_j = x, X_j = y) \), and \( \mathcal{E}(X_jX_k) = \sum_x \sum_y xyP(X_j = x, X_k = y) \). Finally, \( \sigma^2_{Xj} = \mathcal{E}(X_j^2) - [\mathcal{E}(X)]^2 \), and \( \sigma_{jk} = \mathcal{E}(X_jX_k) - \mathcal{E}(X_j)\mathcal{E}(X_k) \).

### 3.2 Data Generation

Samples of \( N \) pairs of latent variable values \( \theta_1 \) and \( \theta_2 \) were drawn from a bivariate normal distribution. The score for person \( i \) on item \( j \) was computed as follows. First, using Eq. (22), \( P(X_j \geq x|\theta_{1i}, \theta_{2i}) \) was computed for \( x = 1, \ldots, 4 \). Second, let \( I \) be an indicator function, and let \( w_{ji} \) be a random number between 0 and 1; then \( X_{ji} = \sum_x I[P(X_j \geq x|\theta_{1i}, \theta_{2i}) > w] \). The resulting item scores are discrete and follow a multinomial distribution.

### 3.3 Design

The between-subject factors were (1) correlation \( \rho_{\theta_1\theta_2} \) (values 0.30, 0.65, and 1), (2) number of items \( J \) (values 5, 10, and 15), and (3) sample size \( N \) (values 50, 250, 500, 750, and 1000). The full factorial design had \( 3 \times 3 \times 5 = 45 \) cells. Each cell was replicated 5000 times. Note that the item scores are unidimensional if \( \rho_{\theta_1\theta_2} = 1 \) and two-dimensional if \( \rho_{\theta_1\theta_2} < 1 \). For each sample, the \( \lambda \)'s were estimated, and the GLB was estimated using function glb.algebraic from the psych r-package (Revelle 2015).

The dependent variables were (1) discrepancy (the difference between the population value of the reliability method and the population reliability (e.g., \( \lambda_4 - \rho_{XX'} \)), (2) bias (the difference between the mean of the sample estimates (e.g., sample estimate denoted \( \hat{\lambda}_4 \), mean denoted \( \bar{\lambda}_4 \)) and the population value (e.g., \( \lambda_4 \) (bias equals \( \bar{\lambda}_4 - \lambda_4 \)), (3) standard deviation of the coefficients (e.g., \( SD(\hat{\lambda}_4) \)), and (4) reliability overestimation (in each design cell, the proportion out of 5000 replicated sample values exceeding \( \rho_{XX'} \) [e.g., \( P(\hat{\lambda}_4 > \rho_{XX'}) \)]).

### 4 Results

#### 4.1 Discrepancy

Table 1 provides the positive discrepancies for the lower bounds. The GLB had negligible discrepancy. Second best methods were \( \lambda_2 \) and \( \lambda_4 \). Methods \( \lambda_5 \) and \( \lambda_6 \) had the largest discrepancy. Except for the GLB, as test length increased, dis-
crepancy decreased, and as correlation between the two latent variables increased, discrepancy increased. Based on discrepancy alone, methods $\lambda_5$ and $\lambda_6$ probably will overestimate reliability not as often as the other methods.

### 4.2 Bias and Standard Deviation

Bias is interesting when it is positive, discrepancy is small, and SD is large. This combination of quantities produces large proportions of reliability overestimates. Tables 2, 3, and 4 show that sample size, more than test length and dimensionality, affects bias and SD; both decrease as $N$ increases. The five lower bounds differ little with respect to SD, so that we will concentrate on bias.

Method $\lambda_5$ and benchmark method $\lambda_2$ had small negative bias, ranging across the design from 0.000 to $-0.016$ ($\lambda_5$) and from $-0.007$ to $-0.018$ ($\lambda_2$). Method $\lambda_6$ had bias ranging from positive when $N = 50$ (0.001 to 0.033) to negative when $N \geq 250$ ($-0.002$ to $-0.020$). Given that method $\lambda_6$ had large discrepancy, we expect the proportion of reliability overestimates to be large for $N = 50$ and small for larger $N$. Bias for methods $\lambda_4$ and GLB was largest and almost always positive, in particular when $J = 10, 15$. In combination with discrepancy that was almost always near 0 or equal to 0, for $\lambda_4$ and the GLB, one may expect large proportions of reliability overestimates.

### 4.3 Reliability Overestimation

For method $\lambda_5$, except when $J = 5$ and $N = 50$, overestimation was negligible (Table 5). For method $\lambda_6$, overestimation was always problematic for $N = 50$ but not for larger $N$. For benchmark $\lambda_2$, for unidimensionality and $N = 50$, proportions were approximately 0.4 but decreased as $N$ increased and also decreased to 0 as
Table 2  Bias and SD of $\lambda_4$, $\lambda_5$, $\lambda_6$, GLB, and $\lambda_2$, for correlation $\rho_{\theta_1\theta_2} = 1$, as a function of test length and sample size

<table>
<thead>
<tr>
<th>J</th>
<th>Method</th>
<th>Bias</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>250</td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_4$</td>
<td>51</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>-9</td>
<td>-11</td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>1</td>
<td>-15</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>40</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>-16</td>
<td>-17</td>
</tr>
</tbody>
</table>

Entries in thousandths; for example, read $-9$ as $-0.009$. Read a blank as 0

Table 3  Bias and SD of $\lambda_4$, $\lambda_5$, $\lambda_6$, GLB, and $\lambda_2$, for correlation $\rho_{\theta_1\theta_2} = 0.65$, as a function of test length and sample size

<table>
<thead>
<tr>
<th>J</th>
<th>Method</th>
<th>Bias</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>250</td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_4$</td>
<td>47</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>-2</td>
<td>-13</td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>1</td>
<td>-16</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>33</td>
<td>-6</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>-16</td>
<td>-18</td>
</tr>
</tbody>
</table>

Entries in thousandths; for example, read $-7$ as $-0.007$. Read a blank as 0
### Table 4  Bias and SD of $\lambda_4$, $\lambda_5$, $\lambda_6$, GLB, and $\lambda_2$, for correlation $\rho_{i,j} = 0.30$, as a function of test length and sample size

<table>
<thead>
<tr>
<th>$J$</th>
<th>Method</th>
<th>Bias</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>250</td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_4$</td>
<td>52</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>$-4$</td>
<td>$-14$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>2</td>
<td>$-17$</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>$-13$</td>
<td>$-18$</td>
</tr>
<tr>
<td>10</td>
<td>$\lambda_4$</td>
<td>63</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>$-3$</td>
<td>$-9$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>25</td>
<td>$-7$</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>75</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>$-9$</td>
<td>$-12$</td>
</tr>
<tr>
<td>15</td>
<td>$\lambda_4$</td>
<td>68</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>$-5$</td>
<td>$-7$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>33</td>
<td>$-2$</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>76</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>$-8$</td>
<td>$-9$</td>
</tr>
</tbody>
</table>

Entries in thousandths; for example, read $-16$ as $-0.016$. Read a blank as 0

### Table 5  Proportions of estimates of $\lambda_4$, $\lambda_5$, $\lambda_6$, GLB, and $\lambda_2$ overestimating $\rho_{XX}$, for correlation $\rho_{i,j} = 1$, as function of test length and sample size

<table>
<thead>
<tr>
<th>$J$</th>
<th>Method</th>
<th>$\rho_{i,j} = 1$</th>
<th>$\rho_{i,j} = 0.65$</th>
<th>$\rho_{i,j} = 0.30$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>250</td>
<td>500</td>
</tr>
<tr>
<td>5</td>
<td>$\lambda_4$</td>
<td>74</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>31</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>24</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>82</td>
<td>64</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>41</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>$\lambda_4$</td>
<td>98</td>
<td>94</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>59</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>99</td>
<td>97</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>42</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>$\lambda_4$</td>
<td>100</td>
<td>99</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>$\lambda_5$</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_6$</td>
<td>85</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>GLB</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>$\lambda_2$</td>
<td>39</td>
<td>20</td>
<td>10</td>
</tr>
</tbody>
</table>

Entries in hundredths; for example, read 31 as 0.31. Read a blank as 0

Note. The reliability was (left to right, top to bottom) 0.761, 0.725, 0.678, 0.865, 0.841, 0.809, 0.906, 0.889, and 0.864
\( \rho_{\theta_1, \theta_2} \) decreased. For method \( \lambda_4 \) and the GLB, irrespective of \( N \), when \( J = 10, 15 \), proportions varied between 0.78 and 1.00. Dimensionality only had little effect on proportions; they were invariably high.

5 Discussion

Discrepancy, bias, and standard deviation together determine proportion of overestimation, but exact numerical results had to be computed using a simulation study. Table 6 provides qualifications of the results based on Tables 1, 2, 3, 4, and 5 and enables us to summarize each of the methods’ results and draw conclusions with respect to their practical usefulness.

For method \( \lambda_4 \), discrepancy is small, but bias is substantial to large, and, moreover, it is positive, thus driving estimates of \( \lambda_4 \) to overestimate \( \rho_{XX'} \); this happens often, and proportion of overestimation is large. The GLB is closer to \( \rho_{XX'} \) and has the same statistical properties as \( \lambda_4 \), hence producing many gross overestimates of \( \rho_{XX'} \). Method \( \lambda_4 \) and GLB both suffer greatly from their tendency to capitalize on chance. This renders their application to moderate sample sizes questionable.

Methods \( \lambda_5 \) and \( \lambda_6 \) have a large discrepancy, whereas the former method has small bias irrespective of sample size \( N \), and the latter method has substantial bias (small \( N \)) to small bias (moderate \( N \)). Combined with a standard deviation that is small enough to have little effect on overestimation if \( N > 250 \), this combination of properties causes methods \( \lambda_5 \) and \( \lambda_6 \) to rarely overestimate \( \rho_{XX'} \). Their large discrepancy speaks to their disadvantage as practical estimates of \( \rho_{XX'} \).

For benchmark method \( \lambda_2 \), Table 6 suggests that discrepancy is small, bias is small irrespective of sample size, and variance did not differ notably between \( \lambda_2 \) and other reliability estimation methods. The magnitude of overestimation usually is small, but for unidimensional data, overestimation may be larger due to \( \lambda_2 \) having small discrepancy.

Compared to method \( \lambda_2 \), methods \( \lambda_4 \), \( \lambda_5 \), and \( \lambda_6 \) and the GLB all seem to underperform. We only studied small samples; hence, a study of the large-sample performance of the methods may be useful. Chance capitalization caused by realistic numbers of items seems to be a principled problem and not easily fixed. Given that

<table>
<thead>
<tr>
<th>Method</th>
<th>Discrepancy</th>
<th>Bias ( N = 50 )</th>
<th>Bias ( N &gt; 250 )</th>
<th>Variance</th>
<th>Overestimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_2 )</td>
<td>Small</td>
<td>Small</td>
<td>Small</td>
<td>No effect</td>
<td>Variable</td>
</tr>
<tr>
<td>( \lambda_4 )</td>
<td>Small</td>
<td>Large</td>
<td>Substantial</td>
<td>No effect</td>
<td>Large</td>
</tr>
<tr>
<td>( \lambda_5 )</td>
<td>Large</td>
<td>Small</td>
<td>Small</td>
<td>No effect</td>
<td>Small</td>
</tr>
<tr>
<td>( \lambda_6 )</td>
<td>Large</td>
<td>Substantial</td>
<td>Small</td>
<td>No effect</td>
<td>Small</td>
</tr>
<tr>
<td>GLB</td>
<td>Negligible</td>
<td>Large</td>
<td>Substantial</td>
<td>No effect</td>
<td>Large</td>
</tr>
</tbody>
</table>
one does not want to report a reliability boosted by chance, reporting a lower bound that does not capitalize on chance, such as $\lambda_2$, may be recommendable. Even method $\lambda_3$ (coefficient $\alpha$), not studied here, may be considered. Such recommendations require further study. In addition, future studies are required to investigate the degree to which the results in this study generalize to continuous item scores. It can be expected that the degree of discrepancy, bias, and precision is slightly different for continuous item scores and for the discrete item scores considered in this study. Methods from the factor analysis approach (Bollen 1989; McDonald 1999) probably also suffer from chance capitalization and require the same kind of evaluation as the methods studied in this article.

References


R.J. De Ayala, The Theory and Practice of Item Response Theory (Guilford Press, New York, 2009)


F.M. Lord, M.R. Novick, Statistical Theories of Mental Test Scores (Addison-Wesley, Reading, 1968)

R.P. McDonald, Test Theory: A Unified Treatment (Erlbaum, Mahwah, 1999)


