Overestimation of reliability by Guttman’s λ4, λ5, and λ6, and the greatest lower bound

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

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

Publication date
2017

Document Version
Final published version

Published in
Quantitative psychology

Citation for published version (APA):
Oosterwijk, P. R., van der Ark, L. A., & Sijtsma, K. (2017). Overestimation of reliability by Guttman’s λ_4, λ_5, and λ_6, and the greatest lower bound. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W-C. Wang (Eds.), *Quantitative psychology: The 81st Annual Meeting of the Psychometric Society, Asheville, North Carolina, 2016* (pp. 159-172). (Springer Proceedings in Mathematics & Statistics; Vol. 196). Springer. https://doi.org/10.1007/978-3-319-56294-0_15

Overestimation of Reliability by Guttman's λ_4, λ_5, and λ_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 λ_4, λ_5, and λ_6 and the greatest lower bound to the reliability (GLB). Method λ_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 λ_4 and the GLB often overestimated test-score reliability. When sample size exceeded 250 observations, methods λ_2, λ_5, and λ_6 provided reasonable to good statistical results, in particular when data were two-dimensional. Benchmark method λ_2 produced the best results.

Keywords Chance capitalization of reliability optimization methods • Greatest lower bound to the reliability • Guttman's λ_2 • Guttman's λ_4 • Guttman's λ_5 • Guttman's λ_6 • Reliability overestimation

P.R. Oosterwijk
Court of Audit, Lange Voorhout 8, The Hague, The Netherlands
e-mail: P.R.Oosterwijk@gmail.com

L.A. van der Ark (✉)
Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15776, 1001 NG, Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

K. Sijtsma
Tilburg School of Social and Behavioral Sciences, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands
e-mail: K.Sijtsma@TilburgUniversity.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 196, DOI 10.1007/978-3-319-56294-0_15

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 λ_4, λ_5, and λ_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 α (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 α 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 λ_4, λ_5, and λ_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 λ_4, λ_5, and λ_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

$$X = T + E. \tag{1}$$

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

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

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

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2, \tag{3}$$

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

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

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

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

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

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

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 λ_4, λ_5, and λ_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 σ_{jk} be the inter-item covariance. One can derive that $\sigma_{T_jT_k} = \sigma_{jk}$, and by definition $\sigma_{E_jE_k} = 0, j \neq k$. If Eq. (3) is rewritten for individual items, we have

$$\sigma^2_{X_j} = \sigma^2_{T_j} + \sigma^2_{E_j}. \tag{7}$$

Covariance matrices $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_T$ are symmetrical and have order $J \times J$, and $\mathbf{\Sigma}_E$ is diagonal and has order $J \times J$, so that

$$\mathbf{\Sigma}_X = \mathbf{\Sigma}_T + \mathbf{\Sigma}_E. \tag{8}$$

Matrix $\mathbf{\Sigma}_X$ is positive definite (pd); that is, for any vector $\mathbf{u}$ of size J, we have $\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u} > 0$ (i.e., $\mathbf{\Sigma}_X$ has a positive determinant); $\mathbf{\Sigma}_T$ and $\mathbf{\Sigma}_E$ are positive semi-definite (psd), so that $\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u} \geq 0$ (i.e., determinants are nonnegative).

The derivation of the methods λ_4, λ_5, and λ_6 and also benchmark λ_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 $\boldsymbol{\lambda_4}$, $\boldsymbol{\lambda_5}$, and $\boldsymbol{\lambda_6}$, GLB, and Benchmark $\boldsymbol{\lambda_2}$*

2.2.1 Method $\boldsymbol{\lambda_4}$

Method λ_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 $\sum_{j=1}^{J} \sigma^2_{E_j}$ in Eq. (6) and consequently a lower bound for $\rho_{XX'}$. Here, we define the upper bound for $\sum_{j=1}^{J} \sigma^2_{E_j}$ typical of method λ_4. Let $\mathbf{u}$ only have elements equal to either $+1$ or -1 so that $\mathbf{u}$ selects items in either of the two test parts of a particular test split. It can be shown that

$$\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u} = \mathbf{u}'\mathbf{\Sigma}_T\mathbf{u} + \mathbf{u}'\mathbf{\Sigma}_E\mathbf{u}, \tag{9}$$

from which it follows that

$$\mathbf{u}'\mathbf{\Sigma}_E\mathbf{u} = \sum_{j=1}^{J} \sigma^2_{E_j} \leq \mathbf{u}'\mathbf{\Sigma}_X\mathbf{u}. \tag{10}$$

The right-hand side of Eq. (10) provides an upper bound for $\sum_{j=1}^{J} \sigma^2_{E_j}$, and method λ_4 finds the vector $\mathbf{u}$ that minimizes $\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u}$, so that

$$\lambda_4 = \max_{\mathbf{u}} \left(1 - \frac{\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u}}{\sigma^2_X}\right). \tag{11}$$

Because $\mathbf{u}$ and $-\mathbf{u}$ provide the same value for $\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u}$, and because vectors $\mathbf{u}$ containing only $+1$s or only -1s do not refer to a test split, $2^{J-1} - 2$ vectors and corresponding products $\mathbf{u}'\mathbf{\Sigma}_X\mathbf{u}$ remain to find λ_4. For test length $J < 20$, one can try all vectors $\mathbf{u}$ within reasonable computing time, and for test length $J \geq 20$, we refer to a procedure proposed by Benton (2015).

2.2.2 Method $\boldsymbol{\lambda_5}$

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

$$\sum_{j=1}^{J} \sigma^2_{T_j} \geq 2\left(\sum_{k \neq j} \sigma^2_{jk}\right)^{\frac{1}{2}}. \tag{12}$$

From Eq. (7) it follows that

$$\sum_{j=1}^{J} \sigma^2_{X_j} = \sum_{j=1}^{J} \sigma^2_{T_j} + \sum_{j=1}^{J} \sigma^2_{E_j}, \tag{13}$$

which implies

$$\sum_{j=1}^{J} \sigma^2_{E_j} \leq \sum_{j=1}^{J} \sigma^2_{X_j} - 2\left(\sum_{k \neq j} \sigma^2_{jk}\right)^{\frac{1}{2}}. \tag{14}$$

The right-hand side of Eq. (14) provides another upper bound for $\sum_{j=1}^{J} \sigma^2_{E_j}$, and one is free to choose the column k that minimizes this upper bound, hence finds a lower bound for $\rho_{XX'}$. To find λ_5, let k vary across each of the J columns of $\mathbf{\Sigma}_T$ and define

$$\lambda_5 = 1 - \frac{\sum_{j=1}^{J} \sigma^2_{X_j} - \max_k \left[2\left(\sum_{k \neq j} \sigma^2_{jk}\right)^{\frac{1}{2}}\right]}{\sigma^2_X}. \tag{15}$$

2.2.3 Method $\boldsymbol{\lambda_6}$

Method λ_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, $\sigma^2_{\epsilon_j}$, is an upper bound to the measurement error variance for item j, $\sigma^2_{E_j}$; that is, $\sigma^2_{E_j} \leq \sigma^2_{\epsilon_j}$, and adding across the J items, we obtain

$$\sum_{j=1}^{J} \sigma^2_{E_j} \leq \sum_{j=1}^{J} \sigma^2_{\epsilon_j} \tag{16}$$

(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,

$$\lambda_6 = 1 - \frac{\sum_{j=1}^{J} \sigma^2_{\epsilon_j}}{\sigma^2_X}. \tag{17}$$

For estimation of λ_6 using covariance matrix $\mathbf{\Sigma}_X$, see Jackson and Agunwamba (1977) and Oosterwijk et al. (2016).

2.2.4 Greatest Lower Bound

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

$$\rho_{XX'} = 1 - \frac{tr(\mathbf{\Sigma}_E)}{\sigma^2_X}, \tag{18}$$

and the GLB is obtained by replacing $tr(\mathbf{\Sigma}_E)$ with $tr(\widetilde{\mathbf{\Sigma}}_E)$, so that

$$GLB = 1 - \frac{tr(\widetilde{\mathbf{\Sigma}}_E)}{\sigma^2_X}. \tag{19}$$

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 λ_2

Guttman (1945) derived three methods, denoted λ_1, λ_2, and λ_3 (equal to coefficient α), 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 λ_4, λ_5, and λ_6 and GLB. The relationship between the three methods and the reliability is

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

In general, λ_1 is considered practically useless, and Sijtsma (2009) and Oosterwijk et al. (2016) have recommended using λ_2 rather than λ_3. Hence, we used λ_2 as benchmark for methods λ_4, λ_5, and λ_6 and GLB. Method λ_2 equals

$$\lambda_2 = 1 - \frac{\sum_{j=1}^{J} \sigma_{X_j}^2 - \left(\frac{J}{J-1} \sum \sum_{j \neq k} \sigma_{jk}^2 \right)^{\frac{1}{2}}}{\sigma_X^2}. \tag{21}$$

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 Sočan 2004). Under particular conditions, method λ_4 has values similar to the GLB (Jackson and Agunwamba 1977; Ten Berge and Sočan 2004). Hence, method λ_4 has almost the same bias relative to $\rho_{XX'}$ as the GLB; Benton (2015) found that method λ_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 λ_4 and the GLB but also for λ_5 and λ_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 θ, item location parameters β_{jx}, and item discrimination parameters α_j, such that

$$P(X_j \geq x | \theta_1, \theta_2) = \frac{\exp[\alpha_{j1}(\theta_1 - \beta_{jx}) + \alpha_{j2}(\theta_2 - \beta_{jx})]}{1 + \exp[\alpha_{j1}(\theta_1 - \beta_{jx}) + \alpha_{j2}(\theta_2 - \beta_{jx})]}. \tag{22}$$

Latent variables θ_1 and θ_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 β_{jx} consisted of an item-specific part τ_j and a category-specific part κ_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 β_{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 β values, which are readily computed.

Discrimination parameters α_j differed across latent variables. For θ_1, $\alpha_{j1} = 1.6$ for the odd-numbered items and $\alpha_{j1} = 0$ for the even-numbered items. For θ_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 $\mathbf{\Sigma}_T$, so as to obtain the numerator of $\rho_{XX'}$ [Eq. (5)]. Because $\sigma_{T_jT_k} = \sigma_{jk}$, we only require the item true-score variances, $\sigma^2_{T_j}$. It also may be noted that $\mathcal{E}(T_j) = \mathcal{E}(X_j)$. First, for fixed values of θ_1 and θ_2, the true score of item j equals

$$T_j|\theta_1, \theta_2 = \sum_x P(X_j \geq x|\theta_1, \theta_2). \tag{23}$$

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

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

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

$$P(X_j = x|\theta_1, \theta_2) = P(X_j \geq x|\theta_1, \theta_2) - P(X_j \geq x + 1|\theta_1, \theta_2). \tag{25}$$

Equation (25) was used to compute covariance matrix $\mathbf{\Sigma}_X$, so as to obtain the denominator of $\rho_{XX'}$ [Eq. (5)], methods λ_4, λ_5, and λ_6, the *GLB*, and benchmark method λ_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, $\sigma^2_{X_j} = \mathcal{E}(X_j^2) - [\epsilon(X)]^2$, 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 θ_1 and θ_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 λs 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 $\bar{\hat{\lambda}}_4$) and the population value (e.g., λ_4) (bias equals $\bar{\hat{\lambda}}_4 - \lambda_4$)), (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 λ_2 and λ_4. Methods λ_5 and λ_6 had the largest discrepancy. Except for the GLB, as test length increased, dis-

Table 1 Discrepancy of λ_4,λ_5, λ_6, GLB, and λ_2, as a function of correlation between latent variables and test length

$\rho_{\theta_1\theta_2}$	J	$\rho_{XX'}$	λ_4	λ_5	λ_6	GLB	λ_2
1	5	0.761	−26	−29	−45		−2
	10	0.865		−34	−14		−1
	15	0.906	−2	−30	−7		−1
0.65	5	0.725	−30	−58	−79	−2	−38
	10	0.841		−49	−26		−18
	15	0.889	−4	−41	−14		−13
0.30	5	0.678	−36	−87	−106	−4	−72
	10	0.809		−66	−35		−36
	15	0.864	−4	−53	−17		−24

Entries for λ_4,λ_5, λ_6, GLB, and λ_2 in thousandths; for example, read −26 as −0.026. Read a blank as 0

crepancy decreased, and as correlation between the two latent variables increased, discrepancy increased. Based on discrepancy alone, methods λ_5 and λ_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 λ_5 and benchmark method λ_2 had small negative bias, ranging across the design from 0.000 to −0.016 (λ_5) and from −0.007 to −0.018 (λ_2). Method λ_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 λ_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 λ_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 λ_4 and the GLB, one may expect large proportions of reliability overestimates.

4.3 Reliability Overestimation

For method λ_5, except when $J = 5$ and $N = 50$, overestimation was negligible (Table 5). For method λ_6, overestimation was always problematic for $N = 50$ but not for larger N. For benchmark λ_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 λ_4, λ_5, λ_6, GLB, and λ_2, for correlation $\rho_{\theta_1\theta_2} = 1$, as a function of test length and sample size

		Bias					SD				
J	Method	50	250	500	750	1000	50	250	500	750	1000
5	λ_4	51	13	4		−2	49	25	18	15	13
	λ_5		−9	−11	−12	−13	54	25	17	14	12
	λ_6	1	−15	−16	−16	−17	62	29	20	16	14
	GLB	40	7		−3	−5	49	25	18	15	13
	λ_2	−16	−17	−16	−16	−16	57	26	18	14	13
10	λ_4	51	19	11	7	5	20	12	9	8	7
	λ_5		−5	−7	−7	−8	29	13	9	7	6
	λ_6	17	−5	−8	−9	−9	29	14	10	8	7
	GLB	58	22	13	8	6	19	11	9	7	6
	λ_2	−9	−10	−10	−10	−10	30	13	9	8	7
15	λ_4	50	21	13	9	7	11	7	6	5	4
	λ_5	−2	−4	−5	−5	−6	20	9	6	5	4
	λ_6	23	−2	−5	−6	−6	17	9	7	5	5
	GLB	55	23	15	11	8	10	7	6	5	4
	λ_2	−7	−7	−7	−7	−7	20	9	6	5	4

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

Table 3 Bias and SD of λ_4, λ_5, λ_6, GLB, and λ_2, for correlation $\rho_{\theta_1\theta_2} = 0.65$, as a function of test length and sample size

		Bias					SD				
J	Method	50	250	500	750	1000	50	250	500	750	1000
5	λ_4	47	9		−4	−5	60	31	22	18	16
	λ_5	−2	−13	−14	−15	−15	68	31	21	18	16
	λ_6	1	−16	−17	−18	−18	76	35	24	19	17
	GLB	33		−6	−9	−10	60	30	21	17	15
	λ_2	−16	−18	−17	−17	−17	71	33	22	18	16
10	λ_4	54	19	10	6	3	25	15	11	9	8
	λ_5	−2	−7	−9	−9	−10	38	17	12	10	8
	λ_6	21	−6	−10	−11	−11	36	18	13	10	9
	GLB	63	22	12	7	5	23	14	10	9	8
	λ_2	−9	−11	−11	−11	−12	39	17	12	10	8
15	λ_4	57	23	13	9	7	13	09	7	6	5
	λ_5	−3	−6	−7	−7	−7	26	11	8	7	6
	λ_6	28	−2	−6	−7	−8	21	11	8	7	6
	GLB	63	26	16	11	9	12	9	7	6	5
	λ_2	−8	−8	−9	−9	−9	26	12	8	7	6

Entries in thousandths; for example, read −7 as −0.007. Read a blank as 0

Table 4 Bias and SD of λ_4, λ_5, λ_6, GLB, and λ_2, for correlation $\rho_{\theta_1\theta_2} = 0.30$, as a function of test length and sample size

		Bias					SD				
J	Method	50	250	500	750	1000	50	250	500	750	1000
5	λ_4	52	11	2	−3	−5	73	36	25	21	19
	λ_5	−4	−14	−15	−16	−16	82	38	26	22	19
	λ_6	2	−17	−18	−19	−20	91	41	28	23	20
	GLB	36	1	−5	−8	−10	72	36	25	21	18
	λ_2	−13	−18	−17	−17	−18	84	39	26	22	19
10	λ_4	63	22	12	7	4	31	18	13	11	9
	λ_5	−3	−9	−10	−11	−11	48	21	15	12	11
	λ_6	25	−7	−11	−13	−13	45	21	15	13	11
	GLB	75	26	14	9	6	29	17	13	11	9
	λ_2	−9	−12	−13	−13	−13	48	21	15	12	10
15	λ_4	68	28	17	12	8	17	11	8	7	6
	λ_5	−5	−7	−8	−8	−9	34	15	10	9	7
	λ_6	33	−2	−7	−8	−9	26	14	10	8	7
	GLB	76	31	19	14	10	15	10	8	7	6
	λ_2	−8	−9	−10	−10	−10	33	15	10	8	7

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

Table 5 Proportions of estimates of λ_4, λ_5, λ_6, GLB, and λ_2 overestimating $\rho_{XX'}$, for correlation $\rho_{\theta_1,\theta_2} = 1$, as function of test length and sample size

		$\rho_{\theta_1,\theta_2} = 1$					$\rho_{\theta_1,\theta_2} = 0.65$					$\rho_{\theta_1,\theta_2} = 0.30$				
J	Method	50	250	500	750	1000	50	250	500	750	1000	50	250	500	750	1000
5	λ_4	74	32	12	4	2	66	25	7	2	1	63	26	9	3	1
	λ_5	31	5				18	1				12				
	λ_6	24	1				14					10				
	GLB	82	64	52	44	37	74	49	37	28	22	71	49	39	28	23
	λ_2	41	25	16	11	8	23	3				14				
10	λ_4	98	94	89	83	78	97	89	83	75	68	96	89	83	74	68
	λ_5	9					6					5				
	λ_6	59	8	100			49	2				45	1			
	GLB	99	97	93	87	84	99	94	88	81	75	98	93	87	80	74
	λ_2	42	22	12	8	5	25	3	1			16				
15	λ_4	100	99	96	91	84	100	98	92	84	71	100	98	93	86	76
	λ_5	2					2					2				
	λ_6	85	19	4	1		77	7				76	7			
	GLB	100	100	99	99	98	100	100	99	97	95	100	100	99	98	95
	λ_2	39	20	10	6	3	22	2				15				

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 λ_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 λ_4, discrepancy is small, but bias is substantial to large, and, moreover, it is positive, thus driving estimates of λ_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 λ_4, hence producing many gross overestimates of $\rho_{XX'}$. Method λ_4 and GLB both suffer greatly from their tendency to capitalize on chance. This renders their application to moderate sample sizes questionable.

Methods λ_5 and λ_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 λ_5 and λ_6 to rarely overestimate $\rho_{XX'}$. Their large discrepancy speaks to their disadvantage as practical estimates of $\rho_{XX'}$.

For benchmark method λ_2, Table 6 suggests that discrepancy is small, bias is small irrespective of sample size, and variance did not differ notably between λ_2 and other reliability estimation methods. The magnitude of overestimation usually is small, but for unidimensional data, overestimation may be larger due to λ_2 having small discrepancy.

Compared to method λ_2, methods λ_4, λ_5, and λ_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 6 Summary of discrepancy, bias, variance, and reliability overestimation

	Discrepancy	Bias $N = 50$	Bias $N > 250$	Variance	Overestimation
λ_2	Small	Small	Small	No effect	Variable
λ_4	Small	Large	Substantial	No effect	Large
λ_5	Large	Small	Small	No effect	Small
λ_6	Large	Substantial	Small	No effect	Small
GLB	Negligible	Large	Substantial	No effect	Large

one does not want to report a reliability boosted by chance, reporting a lower bound that does not capitalize on chance, such as λ_2, may be recommendable. Even method λ_3 (coefficient α), 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

P.M. Bentler, J.A. Woodward, Inequalities among lower bounds to reliability: with applications to test construction and factor analysis. Psychometrika **45**, 249–267 (1980). doi:10.1007/BF02294079

T. Benton, An empirical assessment of Guttman's lambda 4 reliability coefficient, in *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society*, ed. by R.E. Millsap, D.M. Bolt, L.A. van der Ark, W.-C. Wang (Springer, New York, 2015), pp. 301–310

K.A. Bollen, *Structural Equations with Latent Variables* (Wiley, New York, 1989)

L.J. Cronbach, Coefficient alpha and the internal structure of tests. Psychometrika **16**, 297–334 (1951). doi:10.1007/BF02310555

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

L. Guttman, A basis for analyzing test-retest reliability. Psychometrika **10**, 255–282 (1945). doi:10.1007/BF02288892

P.H. Jackson, C.C. Agunwamba, Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I. Algebraic lower bounds. Psychometrika **42**, 567–578 (1977). doi:10.1007/BF02295979

G.F. Kuder, M.W. Richardson, The theory of estimation of test reliability. Psychometrika **2**, 151–160 (1937). doi:10.1007/BF02288391

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)

P.R. Oosterwijk, L.A. Van der Ark, K. Sijtsma, Numerical differences between Guttman's reliability coefficients and the GLB, in *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society*, Beijing, 2015, ed. by L.A. van der Ark, D.M. Bolt, W.-C. Wang, J.A. Douglas, M. Wiberg (Springer, New York, 2016), pp. 155–172

W. Revelle, psych: Procedures for personality and psychological research Version 1.5.8 [computer software] (2015). Evanston, IL. Retrieved from https://cran.r-project.org/web/packages/psych/index.html

A. Shapiro, J.M.F. Ten Berge, The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability. Psychometrika **65**, 413–425. doi:10.1007/BF02296154 2000

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

K. Sijtsma, L.A. Van der Ark, Conceptions of reliability revisited and practical recommendations. Nurs. Res. **64**, 128–136 (2015). doi:10.1097/NNR.0000000000000077

J.M.F. Ten Berge, G. Sočan, The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika **69**, 613–625 (2004). doi:10.1007/BF02289858

J.M.F. Ten Berge, F.E. Zegers, A series of lower bounds to the reliability of a test. Psychometrika **43**, 575–579 (1978). doi:10.1007/BF02293815