A flexible latent class approach to estimating test-score reliability

van der Palm, D.W.; van der Ark, L.A.; Sijtsma, K.

DOI
10.1111/jedm.12053
Publication date
2014
Document Version
Final published version
Published in
Journal of Educational Measurement

JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Winter 2014, Vol. 51, No. 4, pp. 339–357

A Flexible Latent Class Approach to Estimating Test-Score Reliability

Daniël W. van der Palm
Cito, Arnhem
L. Andries van der Ark
University of Amsterdam
Klaas Sijtsma
Tilburg University

The latent class reliability coefficient (LCRC) is improved by using the divisive latent class model instead of the unrestricted latent class model. This results in the divisive latent class reliability coefficient (DLCRC), which unlike LCRC avoids making subjective decisions about the best solution and thus avoids judgment error. A computational study using large numbers of items shows that DLCRC also is faster than LCRC and fast enough for practical purposes. Speed and objectivity render DLCRC superior to LCRC. A decisive feature of DLCRC is that it aims at closely approximating the multivariate distribution of item scores, which might render the method suited when test data are multidimensional. A simulation study focusing on multidimensionality shows that DLCRC in general has little bias relative to the true reliability and is relatively accurate compared to LCRC and classical lower bound methods coefficients α *and* $\lambda 2$ *and the greatest lower bound.*

This study addresses two practical shortcomings of the latent class reliability coefficient (LCRC; Van der Ark, Van der Palm, & Sijtsma, 2011). LCRC is a test-score reliability estimation method based on the unrestricted latent class model (Lazarsfeld, 1950; also see, e.g., Goodman, 1974; Hagenaars & McCutcheon, 2002; McCutcheon, 1987). LCRC produces almost unbiased test-score reliability estimates irrespective of the dimensionality of the data, but it is an inconvenient method due to excessive runtime and the manual labor needed to evaluate the fit of a long series of latent class models containing increasing numbers of latent classes, which easily leads to judgment error. A new version of LCRC called *divisive latent class reliability coefficient* (DLCRC) is proposed that fixes the two problems, thus rendering the method better suited for real-data analysis.

A few remarks about LCRC/DLCRC and data dimensionality are in order. Reise, Waller, and Comrey (2000, p. 294) argued that, unless a test measures a very narrow construct, it is unlikely that test data are unidimensional. Isolating and measuring single attributes is often impossible, and for a group of 6-year olds solving simple addition problems like $3 + 4 = ?$ may already require skills additional to understanding how counting works, such as motor and language skills. True unidimensionality is an ideal and data based on supposedly unidimensional tests will be multidimensional but often close enough to unidimensionality to justify the use of item response theory (IRT; Lord, 1980) models that assume one latent ability or skill. Whether

unidimensionality is realistic or not, however, does not affect LCRC's usefulness; LCRC can be used irrespective of the test's dimensionality.

Other constructs are expressly multidimensional, such as general intelligence, which is composed of an array of near-unidimensional subabilities, for example, designated general and specific aspects of intelligence, verbal and performance intelligence aspects that can be further categorized in more detail, and crystallized and fluid intelligence aspects. A profile of subtest scores may be of interest for diagnostic purposes, but a general intelligence score has to be based on a combination of the subtest scores. In educational measurement, licensure testing, for example, for educators, dentists, medical and nursing professions, and academic proficiency also consist of multiple subtests assessing different abilities and skills necessary for being admitted to practicing a particular occupation. Often, the pass/fail decision is based on a single cut-score and an examinee passing the cut-score shows the minimally required evidence of professional performance for being admitted. Both for general intelligence and licensure testing, LCRC can be used to estimate the composite-score's reliability.

Several reliability estimation methods, such as coefficients α (Cronbach, 1951) $\lambda 2$ (Guttman, 1945), and the greatest lower bound to the reliability (GLB; Bentler & Woodward, 1980; Ten Berge, Snijders, & Zegers, 1981) are equal to test-score reliability when the items or the test parts on which the methods are based are essential τ-equivalent (Lord & Novick, 1968), and reliability methods based on factor analysis provide the best results when congeneric equivalence (Bollen, 1989) holds. Moderate deviations from such assumptions related to unidimensionality may not produce seriously biased statistical outcomes but larger deviations might (Komaroff, 1997; Murphy & DeShon, 2000; Osburn, 2000; Raykov, 1998, 2001; Zimmerman, Zumbo, & Lalonde, 1993). For example, for various methods Van der Ark et al. (2011) found that two-dimensional data produce biased reliability estimates. The authors also found that LCRC provided almost unbiased reliability estimates for unidimensional data and also in a two-dimensional data set.

Some methods that have been proposed to estimate test-score reliability for multidimensional data, such as stratified α (Cronbach, Schoneman, & McKie, 1965) and maximal reliability (Li, Rosenthal, & Rubin, 1996) estimate test-score reliability well (Kamata, Turhan, & Darandari, 2003; Osburn, 2000). These methods require an a priori division of the item set into subsets that each assesses a different knowledge domain, subability or skill. Incorrect item assignment produces underestimation of reliability (Kamata et al., 2003). LCRC does not require a priori item assignment but uses the structure in the multivariate item-score distribution.

In this article, first we discuss the classical test theory framework for test-score reliability and the LCRC method that Van der Ark et al. (2011) followed. Second, we discuss method DLCRC that fixes the two shortcomings of LCRC; introducing and investigating DLCRC was the main goal of this study. Third, an auxiliary goal was using a simulation study to compare LCRC, DLCRC, coefficients α (Cronbach, 1951) and $\lambda 2$ (Guttman, 1945), and the GLB (Bentler & Woodward, 1980;

Ten Berge, Snijders, & Zegers, 1981) with respect to bias and accuracy relative to the test-score reliability. We also estimated the reliability methods in real-data examples. Fourth, we discuss the implications of the results for reliability estimation.

Reliability Theory

Classical test theory (Lord & Novick, 1968) is the theoretical context for this study. Let a test contain J items indexed j with item scores X_j, $j = 1, \ldots, J$. Test score X equals the sum of the J item scores; that is, $X = \sum_{j=1}^{J} X_j$. True score T is the examinee's expected test score over independent repetitions; hence, it is the mean of his propensity distribution (Lord & Novick, 1968, pp. 29–30). The deviation of an examinee's test score from his true score is the random measurement error, E. The classical test model equals $E = X - T$. Measurement error correlates zero with any other variable Y in which it is not included, so that, using ρ to denote the product-moment correlation, in a group of examinees, $\rho_{EY} = 0$. Letting σ^2 denote the variance of a variable, it can be shown that from the assumptions of classical test theory it follows that $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$.

Two tests X and X' are parallel if (1) for each examinee the true scores on the two tests are equal, $T = T'$, and (2) in the group of examinees the tests have equal variance, $\sigma_X^2 = \sigma_{X'}^2$ (Lord & Novick, 1968, p. 48). Parallel tests can be considered a formalization of independent repetitions of a test. Test-score reliability is defined as the product-moment correlation between two parallel tests, and can be shown to equal the proportion of true-score variance on each of the tests,

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_{T'}^2}{\sigma_{X'}^2}. \tag{1}$$

Reliability methods based on factor analysis replace true-score variance by common-factor variance and lead to a definition of reliability, which is smaller than or equal to the classical reliability in Equation 1 (Bentler, 2009). In practice, two parallel tests usually are unavailable and true-score variance is unobservable; consequently, reliability cannot be estimated from real data. Instead, several reliability estimation methods have been proposed that only use one set of data to approximate the correlation between two parallel forms (Equation 1).

We use the framework Van der Ark et al. (2011) proposed and used to analyze dichotomous item scores. Let sample size be denoted N and assume that n_j examinees answered item j correctly and n_{ij} examinees answered both items i and j correctly. Let π_j denote the probability that a randomly drawn examinee answers item j ($j = 1, \ldots, J$) correctly, and let $p_j = n_j/N$ be the sample estimate. Let π_{ij} denote the probability that a randomly drawn examinee answers both items i and j correctly, and let $p_{ij} = n_{ij}/N$ ($i \neq j$) be the sample estimate. For $i = j$, π_{ii} is the probability that a randomly drawn examinee answers item i correctly in two independent repetitions. However, as each item has been administered only once, this probability is unobservable and has to be estimated (Sijtsma & Molenaar, 1987).

For dichotomous items, Molenaar and Sijtsma (1988) showed that reliability (Equation 1) can be written as

$$\rho_{XX'} = \frac{\sum_{i=1}^{J}\sum_{j=1}^{J}\left[\pi_{ij} - \pi_i\pi_j\right]}{\sigma_X^2}. \tag{2}$$

The ratio in (2) can be written as the sum of two ratios,

$$\rho_{XX'} = \frac{\sum\sum_{i\neq j}\left[\pi_{ij} - \pi_i\pi_j\right]}{\sigma_X^2} + \frac{\sum_i\left[\pi_{ii} - \pi_i\pi_i\right]}{\sigma_X^2}. \tag{3}$$

The only unobservable quantity is the joint probability π_{ii}. Van der Ark et al. (2011) showed that methods LCRC, α, and $\lambda 2$ differ only with respect to the way they estimate π_{ii}. Method DLCRC also fits in this framework.

Method LCRC

Van der Ark et al. (2011) proposed to estimate π_{ii} in (3) by means of an unrestricted latent class model. The resulting estimate of $\rho_{XX'}$ is LCRC. We use latent class analysis only as a technical tool to estimate the multivariate item-score distribution. Hence, the latent classes are neither interpreted substantively, nor do they represent a discrete approximation of a latent ability or skill.

Let ξ denote the discrete latent variable and assume ξ has K classes. Conditional on ξ, the manifest item scores are statistically independent. For dichotomous item scores, the latent class model describes the joint probability distribution of the J item scores as,

$$P(X_1 = x_1, \ldots, X_J = x_J) = \sum_{k=1}^{K} P(\xi = k)\prod_{j=1}^{J} P\big(X_j = x_j|\xi = k\big), \tag{4}$$

with $x_j \in \{0, 1\}$. The parameters of the latent class model are the marginal class probabilities, $P(\xi = k)$, and the conditional response probabilities, $P(X_j = x_j|\xi = k)$. For a latent class model with K latent classes (Equation 4), unobservable probability, π_{ii}, equals

$$\pi_{ii} \equiv P(X_i = 1, X_i = 1) = \sum_{k=1}^{K} P(\xi = k)[P(X_i = 1|\xi = k)]^2. \tag{5}$$

LCRC replaces π_{ii} in (3) by the right-hand side of (5). Hence, LCRC equals test-score reliability $\rho_{XX'}$ only if the latent class model with K latent classes perfectly describes multivariate item-score distribution, $P(X_1 = x_1, \ldots, X_J = x_J)$.

Proportion π_{ii} is estimated using the latent class model that has the best relative fit. The fit of different latent class models is compared by means of an *information criterion*: models having a lower information criterion fit better. Consecutive latent class models are estimated, starting with a one-class model, then a two-class model, and so on, until the value of the information criterion no longer decreases. Figure 1 illustrates the strategy of fitting an unrestricted latent class model (left-hand panel). Suppose a three-class model fits the data best, then four models have to be estimated and four information criteria must be computed (in Figure 1 indicated by a

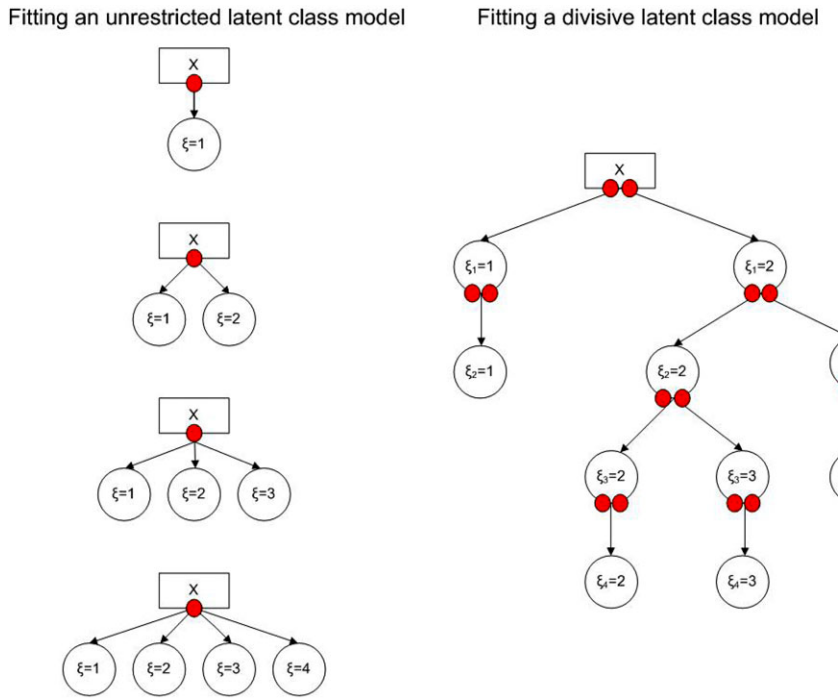

Figure 1. Graphic representation of fitting an unrestricted latent class model (left) and a divisive latent class model (right). A big dot indicates the computation of an information criterion.

big dot). For the one-class, two-class, and three-class models, the information criteria decrease, and for the four-class model the information criterion has increased thus indicating that the three-class model fits best.

Well-known information criteria include the Akaike information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC: (Schwarz, 1978), and AIC3 (Bozdogan, 1987). Simulation studies comparing several information criteria (e.g., Andrews & Currim, 2003; Dias, 2006; Lukociene & Vermunt, 2010) showed that for different types of latent class models, AIC3 retrieved the correct number of classes reasonably well, avoiding both overfitting, which invites chance capitalization, and underfitting, which yields an oversimplified representation of the multivariate item-score distribution. The parameter estimates of the best-fitting model are used to estimate π_{ii} (Equation 5). Probabilities π_j, π_{ij}, and test-score variance σ_X^2 are estimated by means of p_j, p_{ij}, and sample variance S_X^2, respectively.

Method DLCRC

Method LCRC takes much computer time for larger numbers of items and requires considerable manual labor easily inducing judgment error. Method DLCRC solves these problems by using a divisive latent class (DLC) model for estimating π_{ii}. The DLC model describes the multivariate item-score distribution

$P(X_1 = x_1, \ldots, X_J = x_J)$ in terms of (4), but requires K^* rather than K latent classes; typically, $K^* > K$; see Van der Palm (2013, chap. 3) for details. The latent class model estimates the K latent classes simultaneously, whereas the DLC model involves a top-down clustering of examinees into latent classes using one-class and two-class models. Figure 1 (right-hand panel) illustrates the process. First, a one-class and a two-class model are fitted to the entire sample, and an information criterion is computed for both the one-class and the two-class models (in Figure 1, this is indicated by two big dots). If the two-class model fits better than the one-class model, the sample is split into two latent classes. It may be noted that this split is a so-called *soft-partitioning*, which means that each respondent is partly assigned to latent class $\xi_1 = 1$, and partly assigned to latent class $\xi_1 = 2$ (for details, see Van der Palm, 2013, chap. 3). If the two-class model fits better than the one-class model, the sample is not split, the procedure stops, and the one-class solution is accepted as final. If the sample was split into two latent classes (as in Figure 1), in the next steps a one-class and a two-class model are fitted to the subsample in each latent class. If in a latent class the two-class model fits better than the one-class model, the latent class is further split into two latent classes; otherwise, the latent class is accepted as final. The procedure stops if none of the splits improves the fit relative to the local one-class models.

The unrestricted latent class model and the DLC model use the information criteria differently. Each unrestricted latent class model requires a single information criterion (Figure 1, left-hand panel), whereas the DLC-model requires two information criteria for each potential split, one for the one-class model and one for the two-class model (Figure 1, right-hand panel).

Once the DLC model has been estimated, Equation 5 using K^* classes is used to estimate π_{ii}. Probabilities π_j, π_{ij}, and test-score variance σ_X^2 are estimated by means of p_j, p_{ij}, and S_X^2, respectively.

The DLC model is a restricted latent class model. Compared to the unrestricted latent class model this means the following. With each increase in the number of classes, say by one-class, the unrestricted latent class model can optimally divide the sample over all classes, whereas the DLC model only splits the subsample of a single class, leaving the rest of the sample unchanged. As a result, the fit of the unrestricted latent class model improves more by adding an additional latent class than the fit of the DLC model, and the DLC model typically requires more latent classes than the unrestricted latent class model to obtain the same loglikelihood value. Because the DLC model requires more classes, an information criterion such as AIC (see, e.g., Andrews & Currim, 2003; Dias 2006; Lukociene & Vermunt, 2010) that penalizes the model relatively lightly for model complexity so as to allow more classes, may be more appropriate. Based on our simulation studies, we will decide which information criterion yields the best performance.

Other Reliability Estimation Methods

We compared methods LCRC and DLCRC with the lower bounds coefficient α, coefficient $\lambda 2$, and GLB. The three lower bounds are related such that $\alpha \leq \lambda 2 \leq$

GLB $\leq \rho_{XX'}$ (Jackson & Agunwamba, 1977). Let σ_{jk} denote the covariance between items j and k. Coefficient α is defined as

$$\alpha = \frac{J}{J-1}\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sigma_X^2}. \tag{6}$$

Kuder and Richardson (1937) proposed the well-known KR20 equation for binary item scores, and KR20 mathematically equals coefficient α. Coefficient $\lambda 2$ equals

$$\lambda 2 = \frac{\sum\sum_{j\neq k}\sigma_{jk} + \sqrt{\frac{J}{J-1}\sum\sum_{j\neq k}\sigma_{jk}^2}}{\sigma_X^2}. \tag{7}$$

GLB is obtained through a maximization process that finds the largest sum of the item-error variances given the data and the assumptions of classical test theory (e.g., Ten Berge & Sočan, 2004), thus creating the least favorable conditions for the reliability and, therefore, the procedure creates the lower bound of the interval [GLB, 1] in which reliability $\rho_{XX'}$ is located. This is also the greatest value of the set of theoretical lower bounds, including α and $\lambda 2$. GLB tends to overestimate the reliability in small samples and for larger numbers of items (Shapiro & Ten Berge, 2000; Ten Berge & Sočan, 2004).

Method

We performed two studies using artificial data. First, we conducted a simulation study to compare the bias and the accuracy of methods LCRC, DLCRC (both using AIC3 and AIC as an information criterion), α, $\lambda 2$, and GLB. Second, we conducted a runtime study comparing LCRC and DLCRC. Third, we performed a real-data analysis in which we compared the values of the five reliability methods for multidimensional but common educational tests.

Simulation Studies

We used the multidimensional two-parameter logistic model (M2PLM; Reckase, 1997) to generate 0/1 scores. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)$ denote the Q-dimensional latent variable vector; $\boldsymbol{\theta}$ has a Q-variate standard-normal distribution. Let ψ_{jq} denote the discrimination parameter of item j for latent variable q, and let δ_j denote the item location parameter. The M2PLM is defined as,

$$P(X_j = 1|\boldsymbol{\theta}) = \frac{\exp\left[\sum_{q=1}^{Q}\psi_{jq}(\theta_q - \delta_j)\right]}{1+\exp\left[\sum_{q=1}^{Q}\psi_{jq}(\theta_q - \delta_j)\right]}. \tag{8}$$

The M2PLM and $\boldsymbol{\theta}$ were used to compute the population reliability, $\rho_{XX'}$ (Equation 1), and to generate the data. For the computation of $\rho_{XX'}$ by means of Equation 1, the population was defined by 10,000,000 randomly drawn $\boldsymbol{\theta}$s. First, for each θ, and each item, $P(X_j = 1|\boldsymbol{\theta})$ was computed using Equation 8. Second, for each $\boldsymbol{\theta}$ item scores were sampled from $P(X_j = 1|\boldsymbol{\theta})$ and test score X was computed. The

variance of X served as the denominator on the right-hand side of Equation 1. Third, for each $\boldsymbol{\theta}$, true score T was computed as (Lord, 1980, p. 46),

$$T|\boldsymbol{\theta}_n = \sum_{j=1}^{J} P\big(X_j = 1|\boldsymbol{\theta}_n\big)\,. \tag{9}$$

The variance of T served as the numerator in the right-hand side of Equation 1.

The generation of each data set started for each simulee with randomly drawing a latent-variable vector from a Q-variate standard normal distribution, yielding N vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. The correlations r between the Q dimensions were all equal. Subsequently, for each simulee Equation (8) was used to obtain the probability of a particular score for each of the J items, and the item scores were generated by random draws from a multivariate, uniform distribution.

The first simulation study was intended to obtain an impression of bias and accuracy of methods LCRC, DLCRC, α, $\lambda 2$, and GLB. For DLCRC, we investigated whether AIC or AIC3 should be the preferred information criterion. For the sake of completeness, we also investigated AIC and AIC3 for method LCRC. We used a design with main effects instead of a full-factorial design. The design was an extension of the design Van der Ark et al. (2011) used. Van der Ark et al. (2011) studied LCRC (AIC3), α, and $\lambda 2$ but not DLCRC and GLB. Table 1 shows in the upper panel the 14 conditions that are unique to this study and in lower panel the six conditions Van der Ark et al. (2011) studied.

The 15 conditions of this study all concerned multidimensionality. Each item loaded strongly on a primary dimension, and moderately on a secondary and/or a tertiary dimension. The location parameters ranged from easy to difficult for each dimension. Methods LCRC (AIC3), LCRC (AIC), DLCRC (AIC3), DLCRC (AIC), α, $\lambda 2$, and GLB were computed in each of the 14 design conditions, using 1,000 independent replications in each condition.

The first three rows were the reference conditions for short ($J = 6$), long ($J = 18$), and very long ($J = 54$) tests satisfying the M2PLM (Equation 8), both for sample size $N = 1{,}000$, two dimensions ($Q = 2$) that correlated .5 (r), and items that had varying discrimination (ψ_j). A 6-item test length is realistic when items are of the constructed response type requiring the examinee to formulate an answer that involves, for example, the analysis of a political situation or a discussion of a particular social issue, as in history exams (Livingston, 2009). Each of the next 13 conditions differ from the reference conditions by one design factor (i.e., main effect): very long test ($J = 54$); small sample sizes ($N = 50$, 100) to mimic the scale of smaller-classroom testing; latent-variable correlation ($r = .0$, .2) to represent latent variables that are not or weakly associated rather than strong ($r = .5$); three ($Q = 3$) rather than two dimensions to represent dimensional greater complexity; and responding in which guessing plays a part again to mimic classroom testing when multiple-choice items are used. We chose probability .25 to have the answer correct, and adapted Equation 8 as follows. Let the right-hand side be denoted A; then, the M3PLM was defined as $P(X_j|\boldsymbol{\theta}) = .25 + .75A$.

Compared to this study, Van der Ark et al. (2011) concentrated mainly on unidimensionality. They used a standard test condition as reference, defined by

Table 1
Twenty Design Conditions of the Simulation Study

Condition	Design Factors					
	J	N	r	Q	Model	Equal ψ_j
Multidimensional design						
Short	6	1,000	0.5	2	M2PLM	No
Long	18	1,000	0.5	2	M2PLM	No
Very long	54	1,000	0.5	2	M2PLM	No
Short ($N = 50$)	6	50	0.5	2	M2PLM	No
Long ($N = 50$)	18	50	0.5	2	M2PLM	No
Short ($N = 100$)	6	100	0.5	2	M2PLM	No
Long ($N = 100$)	18	100	0.5	2	M2PLM	No
Short ($r = 0.0$)	6	1,000	0.0	2	M2PLM	No
Long ($r = 0.0$)	18	1,000	0.0	2	M2PLM	No
Short ($r = 0.2$)	6	1,000	0.2	2	M2PLM	No
Long ($r = 0.2$)	18	1,000	0.2	2	M2PLM	No
Short (3 dim)	6	1,000	0.5	3	M2PLM	No
Long (3 dim)	18	1,000	0.5	3	M2PLM	No
Short (M3PLM)	6	1,000	0.5	2	M3PLM	No
Long (M3PLM)	18	1,000	0.5	2	M3PLM	No
Design of Van der Ark et al. (2011)						
Standard	6	1,000	–	1	2PLM	No
Polytomous	6	1,000	–	1	GRM	No
Long test	18	1,000	–	1	2PLM	No
Small N	6	200	–	1	2PLM	No
Unequal ψ	6	1,000	–	1	2PLM	Yes
2D-standard	6	1,000	0.0	2	M2PLM	No

six unidimensional items with varying discrimination satisfying the 2PLM, and sample size $N = 1{,}000$. Five additional conditions each differed from the standard design by one design factor: polytomous-item scores ranging from 0 to 4 that were generated using the graded response model (GRM; Samejima, 1969), long test (18 items), small sample size ($N = 200$), unequal discrimination parameters, and two-dimensional data ($Q = 2$) that were generated using the M2PLM. We repeated the Van der Ark et al. (2011) study including DLCRC and GLB.

Table 2 shows the parameter values of the M2PLM for dichotomous items. For 18-item tests, the parameter values of items 7 through 12 and 13 through 18 were identical to those of items 1 through 6, and likewise for the 54-item test the same set of parameter values was assigned to each next item 6-tuple. Van der Ark et al. (2011) used a simple-structure two-dimension condition. In the new conditions, each item loaded on each latent variable. The latent variables were either uncorrelated ($r = .0$), weakly correlated ($r = .2$) or strongly correlated ($r = .5$).

Table 2
Item Parameters of the Multidimensional Two- and Three-Parameter Logistic Models

	Multidimensionality design							Design of Van der Ark et al. (2011)						
	$Q = 2$			$Q = 3$				Standard		Unequal ψ		2D-standard		
Items	ψ_{j1}	ψ_{j2}	δ_j	ψ_{j1}	ψ_{j2}	ψ_{j3}	δ_j	ψ_j	δ_j	ψ_j	δ_j	ψ_{j1}	ψ_{j2}	δ_j
1, 7, 13	2	1	−2	2	1	1	−1.5	1	−2.5	0.5	−2.5	1	0	−2.5
2, 8, 14	2	1	0	2	1	1	1.5	1	−1.5	2	−1.5	1	0	−1.5
3, 9, 15	2	1	2	1	2	1	−1.5	1	−0.5	0.5	−0.5	1	0	−0.5
4, 10, 16	1	2	1.5	1	2	1	1.5	1	0.5	2	0.5	0	1	0.5
5, 11, 17	1	2	0	1	1	2	−1.5	1	1.5	0.5	1.5	0	1	1.5
6, 12, 18	1	2	1.5	1	1	2	1.5	1	2.5	2	2.5	0	1	2.5

Note. For the multidimensional three-parameter logistic model, the guessing parameter equals .25.

Following Van der Ark et al. (2011), the dependent variables were bias and accuracy of LCRC, DLCRC, α, λ2, and GLB. Let r_b denote a reliability estimate in replication b ($b = 1, \ldots, B$; $B = 1{,}000$). Bias was computed as

$$\text{bias} = \frac{1}{B}\sum_{b=1}^{B}(r_b - \rho_{XX'}). \tag{10}$$

For interpretation, Van der Ark et al. (2011) used the following rules of thumb. For absolute bias, $|\text{bias}| < .001$ was considered negligible, $.001 \leq |\text{bias}| < .01$ small, $.01 \leq |\text{bias}| < .02$ medium, $.02 \leq |\text{bias}| < .05$ considerable, and $|\text{bias}| \geq .05$ large. To assess accuracy, the mean absolute error (MAE) was used,

$$\text{MAE} = \frac{1}{B}\sum_{b=1}^{B}|r_b - \rho_{XX'}|. \tag{11}$$

MAE provides the error one can expect for a single data set. MAE$< .002$ was considered negligible, $.002 \leq \text{MAE} < .02$ small, $.02 \leq \text{MAE} < .04$ medium, $.04 \leq \text{MAE} < .10$ considerable, and $\text{MAE} \geq .10$ large.

The second simulation study investigated LCRC and DLCRC runtime. In a single run, we compared LCRC (AIC3) and DLCRC (AIC) runtime for 50, 100, and 500 items. The three-dimensional data were generated using the M2PLM, using latent dimensions that correlated .5 (cf. conditions short (3 dim) and long (3 dim) in Table 1).

Computations in both simulation studies were done using R (R Core Development Team, 2014) and Latent GOLD (Vermunt & Magidson, 2008). All necessary syntax files are available from the first author. DLCRC was estimated using R and Latent GOLD (Vermunt & Magidson, 2008), LCRC, α, and λ2 were estimated using the R package mokken (Van der Ark, 2012), and GLB was estimated using the R package psych (Revelle, 2013).

Real-Data Studies

LCRC, DLCRC, α, λ2, and GLB were computed for eight data sets (Table 3) from educational tests at bachelor-level, administered at the Tilburg School of Social and Behavioral Sciences, The Netherlands. Data dimensionality may affect bias and accuracy of reliability estimates. To assess dimensionality, we used the automated item selection procedure in Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002), which partitions a set of items into unidimensional scales and is available in the R package mokken (Van der Ark, 2012), and a scree-plot of the singular values of the principal axes in multiple correspondence analysis (Greenacre, 2007), which is available in the R package ca (Nenadic & Greenacre, 2007).

Results

Simulation Studies

Table 4 shows the true reliabilities, $\rho_{XX'}$, and the bias of the five reliability estimation methods, and Table 5 shows the accuracy (MAE). The results for LCRC, α, and

Table 3
Real-Data Example; Five Reliability Estimation Methods and Eight Educational Data Sets, Information on Data Dimensionality

Educational Tests			Reliability Estimation Methods					Dimensionality	
Course	*J*	*N*	LCRC	DLCRC	α	λ2	GLB	MSA	CA
Introduction to statistics	20	617	.794	.801	.790	.795	.851	2	1
Experimental methods	30	318	.769	.776	.765	.772	.864	2	1
Construction and analysis of questionnaires	29	306	.736	.743	.730	.740	.846	3	2
Introduction to psychology	38	121	.563	.568	.546	.568	.747	3	4
Social psychology	47	366	.677	.688	.677	.689	.838	4	4
Test Theory	23	248	.567	.577	.560	.581	.753	2	2
Introduction to Mathematics	23	54	.808	.821	.820	.835	.955	2	2
Causal Techniques	24	177	.768	.768	.771	.781	.883	2	2

Note. MSA (Mokken scale analysis): number of scales (i.e., dimensions) in each data set.

λ2 were similar to results Van der Ark et al. (2011) found (Table 4, second panel), suggesting absence of programming errors.

On average, DLCRC (AIC) had the smallest mean absolute bias (Table 4). In particular for smaller *N*, DLCRC (AIC) also had a bias of several hundredths. For DLCRC, using AIC rather than AIC3 produced smaller bias. LCRC (AIC3) and LCRC (AIC) differed little with respect to bias. α and λ2 often showed considerable or large negative bias, in particular when data were multidimensional and item discrimination varied. For other conditions, λ2 had small negative bias and α had small to considerable negative bias. GLB had small to large positive bias in some conditions and small to considerable negative bias in others.

Differences with respect to accuracy were relatively small between reliability methods (Table 5). Compared to Van der Ark et al. (2011), we found smaller differences between methods. In the new design, LCRC and DLCRC were little more accurate than α, λ2, and GLB.

The mean number of latent classes required across the entire simulation study was: LCRC (AIC3): 3.062; LCRC (AIC): 3.412; DLCRC (AIC3): 4.794; and DLCRC (AIC): 5.451. The results illustrate that DLCRC requires more classes than LCRC, and that AIC allows more latent classes than AIC3. Information criteria results for LCRC and DLCRC indicate that AIC has to be preferred for DLCRC, whereas differences are negligible for LCRC suggesting to maintain AIC3 to be consistent with previous work (Van der Ark et al., 2011). In the remainder of this study LCRC (AIC3) and DLCRC (AIC) were used.

Table 4
Bias of Seven Reliability Estimation Methods, Reliability and Bias Values Were Multiplied by 1,000 to Improve Readability

Condition	$\rho_{XX'}$	Bias						
		LCRC AIC3	LCRC AIC	DLCRC AIC3	DLCRC AIC	α	λ2	GLB
				Multidimensional design				
Short	724	−24	−18	−6	−1	−132	−78	−19
Long	887	−8	−7	−10	−9	−48	−21	13
Very long ($J = 54$)	959	−4	−4	−4	−2	−17	−7	12
Short ($N = 50$)	724	−88	−70	−108	−54	−131	−74	12
Long ($N = 50$)	887	−43	−40	−39	−34	−47	−18	62
Short ($N = 100$)	724	−44	−41	−46	−40	−137	−79	−5
Long ($N = 100$)	887	−29	−26	−28	−19	−50	−21	43
Short ($r = 0.0$)	639	−48	−44	−38	−25	−154	−96	−43
Long ($r = 0.0$)	842	−9	−7	−13	−7	−60	−25	17
Short ($r = 0.2$)	679	−42	−38	−25	−16	−146	−89	−34
Long ($r = 0.2$)	864	−10	−8	−12	−10	−55	−24	14
Short (3 dim)	717	−21	−21	−31	−28	−124	−91	−15
Long (3 dim)	884	−10	−9	−12	−6	−45	−32	21
Short (M3PLM)	421	−20	−12	−26	−21	−78	−45	9
Long (M3PLM)	685	−6	−4	−3	−2	−36	−14	47
				Design of Van der Ark et al. (2011)				
Standard	464	−5	1	−9	−7	−16	−7	28
Polytomous	764	−7	−5	−17	1	−14	−7	12
Long test	722	−2	−1	−1	1	−8	−3	46
Small N	464	−5	1	−10	−3	−18	−1	70
Unequal ψ	424	−7	−2	−10	−7	−47	−32	13
2D-standard	315	4	8	−13	1	−77	−46	28

Table 5
Accuracy of Seven Reliability Estimation Methods, Reliability and Accuracy Values (MAE) Were Multiplied by 1,000 to Improve Readability

Condition	$\rho_{XX'}$	MAE						
		LCRC AIC3	LCRC AIC	DLCRC AIC3	DLCRC AIC	α	λ2	GLB
				Multidimensional design				
Short	724	33	33	18	15	132	78	22
Long	887	9	8	10	9	48	21	13
Very long ($J = 54$)	959	4	4	4	2	17	7	12
Short ($N = 50$)	724	98	81	118	65	131	78	54
Long ($N = 50$)	887	44	40	40	35	47	21	62
Short ($N = 100$)	724	52	51	54	51	137	79	41
Long ($N = 100$)	887	29	27	29	21	50	22	43
Short ($r = 0.0$)	639	54	55	42	33	154	96	44
Long ($r = 0.0$)	842	10	9	13	9	60	25	17
Short ($r = 0.2$)	679	45	44	30	23	146	89	35
Long ($r = 0.2$)	864	11	9	12	10	55	24	15
Short (3 dim)	717	27	27	35	33	124	91	24
Long (3 dim)	884	12	11	14	9	45	32	22
Short (M3PLM)	421	32	32	32	31	78	46	25
Long (M3PLM)	685	13	13	12	12	36	16	47
				Design of Van der Ark et al. (2011)				
Standard	464	23	25	22	22	24	21	33
Polytomous	764	11	11	21	10	15	11	14
Long test	722	10	11	11	10	12	10	46
Small N	464	47	48	51	45	47	45	76
Unequal ψ	424	26	28	25	26	48	35	25
2D-standard	315	30	32	38	32	77	48	37

Table 6
Reliability, Estimated Reliability, and Computation Time for LCRC (AIC3) and DLCRC (AIC) for Two- and Three-Dimensional Data, Consisting of 50, 100, and 500 items

		Reliability			Computation Time	
Q	J	$\rho_{XX'}$	LCRC	DLCRC	LCRC	DLCRC
2	50	.955	.950	.951	9 min 39 s	16 s
	100	.977	.976	.977	23 min 5 s	62 s
	500	.995	.995	.995	1 h 39 min 49 s	4 min 19 s
3	50	.975	.970	.971	13 min 37 s	13 s
	100	.987	.985	.985	29 min 8 s	43 s
	500	.997	.995	.995	4 h 4 min 7 s	6 min 43 s

Note. Q = number of dimensions; J = number of items; $\rho_{XX'}$ = reliability. The computation times for α, $\lambda 2$, and GLB were negligibly small and were therefore excluded.

Table 6 shows that computation time of LCRC may be too long for practical purposes. Computation time for DLCRC stays within reasonable limits and is not prohibitive for using longer tests.

Real-Data Results

Mokken scale analysis revealed multidimensionality in each data set (Table 3). Correspondence analysis suggested two data sets were unidimensional. LCRC (AIC3), DLCRC (AIC), α, and $\lambda 2$ showed small differences. DLCRC and $\lambda 2$ often produced somewhat higher estimates than LCRC and α. GLB produced much higher estimates than the other methods but given small sample sizes results likely capitalized on chance. For the *Introduction to Mathematics* and *Causal Techniques* data sets, the difference between DLCRC and $\lambda 2$ was relatively large; $\lambda 2$ was .014 and .013 units higher than DLCRC, respectively. For the other courses, DLCRC and $\lambda 2$ were comparable.

Discussion

Unlike method LCRC, method DLCRC does not suffer from runtime problems and avoids manual labor and the risk of making judgmental errors; hence, DLCRC improves upon LCRC, especially when AIC is used to assess goodness of fit. DLCRC often closely approximates the true reliability, suggesting that speeding up the method using the DLC method improves upon LCRC that uses unrestricted latent class analysis. For multidimensional data, bias differences between DLCRC, α, and $\lambda 2$ were pronounced. DLCRC may be used whenever test data are multidimensional.

Some additional observations are that coefficients α and $\lambda 2$ are seriously biased when data are clearly multidimensional or items have different discrimination parameters. Also, coefficient $\lambda 2$ produces a better reliability estimate than α and GLB. By definition, $\lambda 2$ is closer to the population reliability than α, and GLB produces better estimates when the sample size exceeds 1,000.

Limitations of the simulation study with respect to bias and accuracy are the following. First, as we were primarily interested in the effect of the reliability estimation methods on bias and accuracy we only investigated the main effects of the other independent factors such as test length and number of items. Second, the number of levels of each independent factor was limited to keep the study manageable. We believe our study has captured the most important trends but a fully crossed design including more factor levels may provide additional insights.

The real-data sets varied from 54 to 617 observations, and estimation results were subject to sampling error. True reliability was unknown so that we did not know which reliability estimate was closest to true reliability. Two small data sets ($N = 54$ and 177) yielded $\lambda 2$ higher than DLCRC but the sample size may be prohibitive for drawing conclusions. $\lambda 2$ is a lower bound to the reliability and DLCRC often has little bias; hence, it may be reasonable to report both $\lambda 2$ and DLCRC.

For the estimation of the latent class models used for LCRC and DLCRC, we recommend trying many (approximately 100) different sets of starting values for the parameters (cf. Vermunt, Van Ginkel, Van der Ark, & Sijtsma, 2008). This reduces the probability of ending up in a local maximum, which possibly has a small effect on LCRC and DLCRC. In the R code that is available upon request from the first author, the user can specify the number of different sets of starting values for the (divisive) latent-class model.

Finally, we have used the DLC model only as a density estimation technique; the latent classes only describe the associations in the multivariate distribution and are not intended to have substantive meaning. Exploring whether the DLC model is also useful to model substantive phenomena is a topic for future research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Andrews, R. L., & Currim, I.S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, *40*, 235–243.

Bentler, P. A. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137–143.

Bentler, P. A., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*, 249–267.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *10*, 255–282.

Cronbach, L. J., Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified parallel tests. *Educational and Psychological Measurement*, *25*, 291–312.

Dias, J. G. (2006). Model selection for the binary latent class model: A Monte Carlo simulation. In V. Batagelj, H. H. Bock, A. Ferligoj, & A. Žiberna (Eds.), *Data science and classification* (pp. 91–99). Berlin, Germany: Springer.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Greenacre, M. (2007). *Correspondence analysis in practice*. London, UK: Chapman & Hall/CRC.

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

Hagenaars, J. A. P., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrika*, *42*, 567–578.

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Multidimensional composite scale scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. Retrieved October 6, 2014, from http://www.learningace.com/doc/2646467/5a6eda47494992fc5c0591e99aa942f4/md_rel_paper.pdf

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α. *Applied Psychological Measurement*, *21*, 337–348.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *3*, 151–160.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II. Vol. IV: Measurement and prediction* (pp. 361–412). Princeton, NJ: Princeton University Press.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*, 98–107.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. *Educational Testing Service R & D Connections, No. 11*. Retrieved March 16, 2013, from http://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Lord, F. M. (1980). *Applicatons of item response theory to practical testing problems*. Hilldale, NJ: Lawrence Erlbaum.

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

Lukociene, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241–250). Berlin, Germany: Springer.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden*, *9*(28), 115–126.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.

Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, *20*(3), 1–13.

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, *5*, 343–355.

R Development Core Team. (2014). *R: A language and environment for statistical computing* [computer programming language]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved August 3, 2014, from http://www.R-project.org

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement*, *22*, 375–385.

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297.

Revelle, W. (2013) *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved October 6, 2014, from http://cran.r-project.org/web/packages/psych/index.html

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved October 6, 2014, from http://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

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

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*, 79–97.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Ten Berge, J. M. F., Snijders, T. A. B., & Zegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika*, *46*, 201–213.

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

Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27.

Van der Ark, L. A., Van der Palm, D. W., & Sijtsma, K. (2011). A latent-class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*, 380–392.

Van der Palm, D. W. (2013). *Latent class models for density estimation, with applications in missing data imputation and test-score reliability estimation*. Unpublished doctoral dissertation, Tilburg University, The Netherlands. Retrieved December 3, 2013, from http://www.dvdpalm.nl/thesis.pdf

Vermunt J. K., & Magidson J. (2008). *LG-syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*, 369–397.

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

Authors

DANIËL VAN DER PALM is Researcher at the Psychometrics Department of Cito, Amsterdamseweg 13, 6814 CM, Arnhem, The Netherlands; Daniel.vanderPalm@cito.nl. His primary research interests include missing data, latent class models, reliability analysis, and statistical programming.

ANDRIES VAN DER ARK is Associate Professor of Methodology and Statistics at the Research Institute of Child Development and Education at the University of Amsterdam,

PO Box 15776, 1001 NG, Amsterdam, The Netherlands; L.A.vanderArk@uva.nl. His primary research interests include reliability analysis, nonparametric item response theory and Mokken style analysis, and the development of categorical marginal models for the analysis of test and questionnaire data.

KLAAS SIJTSMA is Professor of Methods of Psychological Research at the Tilburg School of Social and Behavioral Sciences, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands; k.sijtsma@tilburguniversity.edu. His research specializes in psychometrics, in particular all issues related to the measurement of psychological attributes by means of tests and questionnaires.