

## **A Latent Class Approach to Estimating Test-Score Reliability**

L. Andries van der Ark, Daniël W. van der Palm and Klaas Sijtsma  
*Applied Psychological Measurement* published online 9 March 2011  
DOI: 10.1177/0146621610392911

The online version of this article can be found at:  
<http://apm.sagepub.com/content/early/2011/03/09/0146621610392911>

---

Published by:



<http://www.sagepublications.com>

**Additional services and information for *Applied Psychological Measurement* can be found at:**

**Email Alerts:** <http://apm.sagepub.com/cgi/alerts>

**Subscriptions:** <http://apm.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>

**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

# A Latent Class Approach to Estimating Test-Score Reliability

Applied Psychological Measurement  
 XX(X) 1–13  
 © The Author(s) 2011  
 Reprints and permission:  
[sagepub.com/journalsPermissions.nav](http://sagepub.com/journalsPermissions.nav)  
 DOI: 10.1177/0146621610392911  
<http://apm.sagepub.com>



L. Andries van der Ark<sup>1</sup>, Daniël W. van der Palm<sup>1</sup>,  
 and Klaas Sijtsma<sup>1</sup>

## Abstract

This study presents a general framework for single-administration reliability methods, such as Cronbach's alpha, Guttman's lambda-2, and method MS. This general framework was used to derive a new approach to estimating test-score reliability by means of the unrestricted latent class model. This new approach is the latent class reliability coefficient (LCRC). Unlike other single-administration reliability methods, LCRC places few restrictions on the item scores. A simulation study showed that if data are multidimensional or if double monotonicity does not hold, then LCRC is less biased relative to the true reliability than Cronbach's alpha, Guttman's lambda-2, method MS, and the split-half reliability coefficient.

## Keywords

latent class models, reliability, test theory, true score theory

Test-score reliability, denoted  $\rho_{XX'}$ , is one of the most reported statistics in social and behavioral science research. This study adopts the definition proposed by Lord and Novick (1968, p. 61). Let  $X$  be the test score, which is defined as the sum of the  $J$  item scores  $X_j$  ( $j = 1, \dots, J$ ), so that  $X = \sum_{j=1}^J X_j$ . In the population, test score  $X$  has expectation  $\mu_X$  and variance  $\sigma_X^2$ . Let  $T$  be the unobservable true score (Lord & Novick, 1968, chaps. 2 and 3), defined as a testee's expectation of  $X$  across his or her propensity distribution of independent test repetitions. In the population,  $T$  has expectation  $\mu_T$  and variance  $\sigma_T^2$ . Test-score reliability is defined as the product-moment correlation between two sets of independent test scores from two different but interchangeable tests known as parallel tests (which replace two independent repetitions), and equals the ratio of true score and test score variances,

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

<sup>1</sup>Tilburg University, Netherlands

## Corresponding Author:

L. Andries van der Ark, Department of Methodology and Statistics, School of Social and Behavioral Sciences, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, Netherlands  
 Email: [a.vdark@uvt.nl](mailto:a.vdark@uvt.nl)

For reliability estimation one needs sets of test scores collected from parallel tests, or from the same test on two different occasions so that the test is its own parallel test. Because, in practice, two sets of parallel test scores are usually unavailable, researchers often resort to estimating reliability from the item scores obtained in a single test administration using interitem covariances or from the correlation between the scores on two test halves. Unless the item scores are essentially  $\tau$ -equivalent (i.e., a weak form of parallelism; Lord & Novick, 1968, p. 50) or the scores on test halves are parallel, test-score reliability is underestimated. Thus, it is appealing to find single-administration methods that show little bias relative to  $\rho_{XX'}$ . This study proposes such a method.

Reliability methods that focus on the interitem covariances in the test are often called internal consistency methods. Unfortunately, the term *internal consistency* is also used to suggest that a high value produced by such a reliability method means that the items measure the same attribute, as if the test were 1-factorial. This misconception has persisted despite persuasive warnings by, for example, Cortina (1993), Schmitt (1996), and Sijtsma (2009). To avoid misunderstanding, the present study speaks of single-administration reliability instead of internal consistency reliability.

The most frequently used single-administration reliability estimate is Cronbach's alpha (Cronbach, 1951; more than 5,500 citations on Web of Science). Ten Berge and Zegers (1978) showed that Cronbach's alpha is the smallest lower bound in an infinite series of lower bounds to the reliability. These lower bounds are denoted  $\mu_0, \mu_1, \dots$  (with  $\mu_0 = \alpha$ ), and related  $\mu_0 \leq \mu_1 \leq \dots \leq \rho_{XX'}$ . Strict equalities are obtained when the  $J$  items in the test are essentially  $\tau$ -equivalent (Lord & Novick, 1968, p. 50). Because essential  $\tau$ -equivalence is never met in real data, in practice, strict inequalities hold. Ten Berge and Zegers noted that in real data,  $\mu_1$  may improve upon alpha, but that the other  $\mu$ -coefficients usually provide negligible gain. Coefficient  $\mu_1$  equals Guttman's (1945) lambda-2 coefficient. Both alpha and lambda-2 were included in the present study. Note that words rather than symbols have been used when referring to reliability estimates (e.g.,  $\mu_0$  rather than  $\mu_0$ ) to avoid confusion with parameters that use the same symbol (e.g.,  $\mu_T$  is the population mean).

Many different single-administration methods exist, such as Revelle's beta (Revelle, 1979; Zinbarg, Revelle, Yovel, & Li, 2005), the Kristof reliability coefficient (Sedere & Feldt, 1977), and the Feldt coefficient (Sedere & Feldt, 1977). Bentler and Woodward (1980) and Ten Berge, Snijders, and Zegers (1981) solved the problem of finding the greatest lower bound to the reliability. Reliability methods based on structural equation modeling (e.g., Bentler, 2009; Green & Yang, 2009; Raykov, 1997; Raykov & Shrout, 2002) conceptualize a different reliability definition.

Molenaar and Sijtsma (1988; also Sijtsma, 1988; Sijtsma & Molenaar, 1987; Van der Ark, 2010) proposed the single-administration method MS. Method MS is available in the computer package MSP (Molenaar & Sijtsma, 2000) under the name of rho. Sijtsma and Molenaar (1987) simulated binary item scores under the restrictive item response model known as the double monotonicity model (Mokken, 1971, p. 174; for polytomous items, see Molenaar, 1997), and found that method MS and two related methods proposed by Mokken (1971, pp. 142-147) provided almost unbiased estimates of  $\rho_{XX'}$ . The results also suggested that the three estimates were less efficient than alpha and lambda-2. These authors recommended using alpha or lambda-2 if the sample size is small because the other methods may accidentally overestimate  $\rho_{XX'}$ , but for greater sample sizes they recommended method MS. The statistical properties of method MS for polytomously scored items and for item scores generated by less restrictive item response models have not been investigated thus far.

In this study, a new reliability estimation method is presented that does not require restrictive conditions such as essential  $\tau$ -equivalence (coefficients alpha and lambda-2) or the double

monotonicity model (method MS). First, a general framework for single-administration methods is discussed that is based on derivations in Molenaar and Sijtsma (1988). Second, it is proposed to use the latent class model (LCM) to estimate particular parameters needed to estimate the newly proposed reliability method called the *latent class reliability coefficient* (LCRC). It is demonstrated that the LCRC estimates  $\rho_{XX'}$  with negligible bias (unlike alpha and lambda-2) and without relying on a strong model (unlike method MS). Third, the bias and the accuracy of methods alpha, lambda-2, MS, and LCRC are investigated.

### A Framework for Single-Administration Methods

Throughout, it is assumed that all items in the test have the same number of ordered answer categories. This number is denoted  $m + 1$ . The presented framework is also valid for test scores based on items with different numbers of answer categories, but this possibility was ignored here because of the complexity of the presentation and, moreover, because it represents a situation psychometricians often prefer to discourage as it may lead to the differential weighing of items. Notation  $g, h, i,$  and  $j$  is used to index items, and  $x$  and  $y$  to index item scores that run from  $0, 1, \dots, m$ . Let  $\pi_{x(j)} = P(X_j \geq x)$  ( $j = 1, \dots, J; x = 0, \dots, m$ ) be the probability of obtaining at least a score  $x$  on item  $j$ . These probabilities are referred to as marginal cumulative probabilities. It may be noted that  $\pi_{0(j)} = 1$  by definition. Likewise, let  $\pi_{x(i),y(j)} = P(X_i \geq x, X_j \geq y)$  ( $i, j = 1, \dots, J; x, y = 0, \dots, m$ ) be the probability of obtaining at least a score  $x$  on item  $i$  and at least a score  $y$  on item  $j$ . These probabilities are referred to as joint cumulative probabilities.

For  $i = j$ , the joint cumulative probability  $\pi_{x(i),y(i)}$  denotes the probability of obtaining at least score  $x$  and at least score  $y$  on two independent administrations of the same item to the same respondents. This is only possible theoretically because in real life, respondents would remember the second time what they answered the first time, and local independence would be violated. Thus, in practice these independent repetitions are unavailable, and joint cumulative probabilities  $\pi_{x(i),y(i)}$  cannot be estimated using simple bivariate fractions derived from single-administration data; hence, more involved estimation methods are needed.

Molenaar and Sijtsma (1988) showed that reliability (Equation 1) can be written as

$$\rho_{XX'} = \frac{\sum_{i=1}^J \sum_{j=1}^J \sum_{x=1}^m \sum_{y=1}^m [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} \tag{2}$$

Equation 2 can be split into two ratios,

$$\rho_{XX'} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\sum_i \sum_x \sum_y [\pi_{x(i),y(i)} - \pi_{x(i)}\pi_{y(i)}]}{\sigma_X^2} \tag{3}$$

Equation 3 is used as a general framework for single-administration reliability methods. The numerator of the first ratio in Equation 3 can be estimated using the marginal and joint cumulative fractions in the data. This numerator is called the *observable numerator*. It is the sum of  $J(J - 1)m^2$  terms.

The numerator of the second ratio in Equation 3 contains the joint cumulative probabilities pertaining to the same item,  $\pi_{x(i),y(i)}$ . This numerator is called the *unobservable numerator*. It is the sum of  $Jm^2$  terms. The single-administration reliability methods alpha, lambda-2, MS, and LCRC differ only in the way they approximate the unobservable numerator in Equation 3.

## Cronbach's Alpha

Cronbach's alpha can be cast in terms of Equation 3, with each term of the unobservable numerator replaced by the mean of the terms in the observable numerator. Let  $\sigma_{ij}$  denote the covariance between  $X_i$  and  $X_j$ ; then alpha is defined as

$$\text{alpha} = \frac{J}{J-1} \times \frac{\sum_{i \neq j} \sum \sigma_{ij}}{\sigma_X^2}. \quad (4)$$

Because  $\sigma_{ij} = \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]$ , Equation 4 is equivalent to

$$\text{alpha} = \frac{\frac{J}{J-1} \times \sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2}. \quad (5)$$

For any constant  $a$ , one may write  $\frac{J}{J-1} \times a = \frac{J-1+1}{J-1} \times a = \frac{J-1}{J-1} \times a + \frac{1}{J-1} \times a = a + \frac{1}{J-1} \times a$ , and then use this result to split Equation 5 into two parts,

$$\text{alpha} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\frac{1}{J-1} \left\{ \sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}] \right\}}{\sigma_X^2}. \quad (6)$$

Let  $\bar{\pi}$  be the mean of all terms of the observable numerator in Equation 3; that is,

$$\bar{\pi} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{J(J-1)m^2}. \quad (7)$$

It follows from Equation 7 that

$$\begin{aligned} \sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}] &= \bar{\pi}J(J-1)m^2 \Leftrightarrow \\ \frac{1}{J-1} \times \sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}] &= \bar{\pi}Jm^2 \\ &= \sum_i \sum_x \sum_y \bar{\pi}. \end{aligned} \quad (8)$$

Taking Equation 6 and substituting the numerator of the second ratio on the right-hand side by the sum on the right-hand side of Equation 8 yields

$$\text{alpha} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\sum_i \sum_x \sum_y \bar{\pi}}{\sigma_X^2}. \quad (9)$$

Compared to  $\rho_{XX'}$  (Equation 3), in coefficient alpha (Equation 9), each term in the unobservable numerator in Equation 3 has been replaced by the mean of the terms of the observable numerator.

Equations 9 and 3 have been used to explain why Cronbach's alpha is a lower bound to the reliability. In the definition of  $\rho_{XX'}$ , the term  $\sum_x \sum_y [\pi_{x(i),y(i)} - \pi_{x(i)}\pi_{y(i)}]$  (part of the

unobservable numerator in Equation 3) is the covariance between two replications of the same item. In Cronbach’s alpha (Equation 9), this term is estimated by  $\sum_x \sum_y \bar{\pi}$ , which is the mean interitem covariance. It follows from classical test theory that the covariance between two independent replications of the same item is at least as large as the covariance between two different items. Hence, the numerator of the second fraction in Equation 9 cannot exceed the unobservable numerator in Equation 3, and  $\alpha \leq \rho_{XX'}$ .

**Guttman’s Lambda-2**

Like Cronbach’s alpha, Guttman’s (1945) lambda-2 can be cast in terms of Equation 3. Guttman’s lambda-2 is defined as

$$\text{lambda-2} = \frac{\sum_{i \neq j} \sum \sigma_{ij} + \sqrt{\frac{J}{J-1} \sum_{i \neq j} \sum \sigma_{ij}^2}}{\sigma_X^2},$$

and can be written as

$$\text{lambda-2} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\sqrt{\frac{J}{J-1} \sum_{i \neq j} \sum_x \sum_y \left\{ \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}] \right\}^2}}{\sigma_X^2}. \tag{10}$$

Compared to  $\rho_{XX'}$  (Equation 3), in Guttman’s lambda-2 (Equation 10) the unobservable numerator in Equation 3 has been replaced by the square root of a weighted sum of squared sums of terms in the observable numerator. The proof that  $\alpha \leq \text{lambda-2}$  is a standard result in classical test theory (e.g., Ten Berge & Zegers, 1978).

**Method MS**

Method MS was based on the framework represented by Equation 3. Let  $\tilde{\pi}_{x(i),y(i)}$  be an estimator of  $\pi_{x(i),y(i)}$  to be discussed later. Method MS equals Equation 3, in which  $\pi_{x(i),y(j)}$  has been replaced by  $\tilde{\pi}_{x(i),y(j)}$ ; that is,

$$\text{MS} = \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} + \frac{\sum_i \sum_x \sum_y [\tilde{\pi}_{x(i),y(i)} - \pi_{x(i)}\pi_{y(i)}]}{\sigma_X^2}. \tag{11}$$

The procedure for finding estimator  $\tilde{\pi}_{x(i),y(i)}$  is sketched briefly using an artificial example. For detailed descriptions, the present authors refer to Sijtsma and Molenaar (1987) for dichotomously scored items, and to Molenaar and Sijtsma (1988) for polytomously scored items; see Van der Ark (2010) for computational details.

Consider the marginal cumulative probabilities of four items, each with three ordered scores (Table 1). The first step in finding  $\tilde{\pi}_{x(i),y(i)}$  is to rank all informative (i.e., excluding  $\pi_{0(i)} = 1, i = 1, \dots, 4$ ) marginal cumulative probabilities from small to large. For the numerical example, Table 1 shows that this rank order is

$$\pi_{2(4)} < \pi_{2(3)} < \pi_{2(2)} < \pi_{2(1)} < \pi_{1(4)} < \pi_{1(3)} < \pi_{1(2)} < \pi_{1(1)}, \tag{12}$$

**Table 1.** Example of Marginal Cumulative Probabilities

	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$\pi_{0(i)}$	1.00	1.00	1.00	1.00
$\pi_{1(i)}$	.90	.80	.70	.60
$\pi_{2(i)}$	.50	.40	.30	.20

**Table 2.** Marginal Cumulative Probabilities (boldface) and Joint Cumulative Probabilities

	$\pi_{2(4)}$	$\pi_{2(3)}$	$\pi_{2(2)}$	$\pi_{2(1)}$	$\pi_{1(4)}$	$\pi_{1(3)}$	$\pi_{1(2)}$	$\pi_{1(1)}$
	<b>.20</b>	<b>.30</b>	<b>.40</b>	<b>.50</b>	<b>.60</b>	<b>.70</b>	<b>.80</b>	<b>.90</b>
$\pi_{2(4)}$	<b>.20</b>	NA	.10	.10	.20	NA	.20	.20
$\pi_{2(3)}$	<b>.30</b>	.10	NA	.30	.30	NA	.30	.30
$\pi_{2(2)}$	<b>.40</b>	.10	.30	NA	.40	.40	NA	.40
$\pi_{2(1)}$	<b>.50</b>	.20	.30	.40	NA	.50	.50	NA
$\pi_{1(4)}$	<b>.60</b>	NA	.30	.40	.50	NA	.60	.60
$\pi_{1(3)}$	<b>.70</b>	.20	NA	.40	.50	NA	.70	.70
$\pi_{1(2)}$	<b>.80</b>	.20	.30	NA	.50	.60	NA	.80
$\pi_{1(1)}$	<b>.90</b>	.20	.30	.40	NA	.60	.70	NA

but in other examples, different orderings are possible. If ties occur in Equation 12, the marginal cumulative probabilities involved are pooled (see Van der Ark, 2010, for details).

The second step is to create a  $Jm \times Jm$  matrix of joint cumulative probabilities in which the rows and columns are ordered by the corresponding marginal cumulative probabilities, which have been ordered by increasing magnitude (Table 2). In Table 2, NA refers to  $\pi_{x(i),y(i)}$ , the unobservable joint cumulative probability (Equation 3), which is estimated by  $\tilde{\pi}_{x(i),y(i)}$ , for all  $i$  (Equation 11). For matrices of joint cumulative probabilities that are constructed as in Table 2, Mokken (1971, pp. 132-133) proved that if the double monotonicity model holds, then in each row and each column the entries are nondecreasing. Method MS uses this ordering property for estimating the unobservable joint cumulative probabilities by means of linear interpolation. Molenaar and Sijtsma (1988) discussed eight possible linear interpolation methods, each yielding a different estimate for each unobservable joint cumulative probability. For some of the unobservable joint cumulative probabilities (i.e., the NAs in the first and last rows and the first and last columns of Table 2), it is not possible to apply all eight linear interpolation methods, and  $\tilde{\pi}_{x(i),y(i)}$  is estimated as the mean of the available methods.

The assumption that the double monotonicity model holds is rather restrictive because under this model  $\theta$  is unidimensional (unidimensionality), the item scores are independent given  $\theta$  (local independence),  $P(X_i \geq x|\theta)$  is nondecreasing in  $\theta$  for all  $x$  and all  $i$  (monotonicity), and  $P(X_i \geq x|\theta)$  and  $P(X_j \geq y|\theta)$  do not intersect for all  $i \neq j$ . If the double monotonicity model does not hold for the data at hand, then  $\tilde{\pi}_{x(i),y(i)}$  may be a poor approximation to the unobservable joint cumulative probabilities,  $\pi_{x(i),y(i)}$ .

## Latent Class Reliability Coefficient

Like the previously discussed methods, the LCRC is based on the framework represented by Equation 3. As with method MS, the joint cumulative probabilities  $\pi_{x(i),y(i)}$  are approximated assuming a statistical model. For the LCRC, the statistical model is the unconstrained LCM (Hagenaars & McCutcheon, 2002; Lazarsfeld, 1950), which only assumes that the items are

independent given class membership. This is the local independence assumption. Compared to the double monotonicity model underlying method MS, the unconstrained LCM underlying the LCRC is unrestrictive because it does not assume unidimensionality, monotonicity, and nonintersecting item response functions. Therefore, it is expected that the unconstrained LCM describes associations in data well even if properties typically assumed in item response theory, such as unidimensionality, monotonicity, and nonintersecting item response functions, do not hold. This gives the LCRC an advantage over method MS because within the framework of Equation 3, reliability is estimated well if the statistical model approximates the unobserved joint cumulative probabilities  $\pi_{x(i),y(i)}$  well.

For local independence given a discrete latent variable  $\xi$  with  $K$  classes, the unconstrained LCM is defined as

$$P(X_1 = x_1, \dots, X_J = x_J) = \sum_{k=1}^K P(\xi = k) \prod_{j=1}^J P(X_j = x_j | \xi = k). \tag{13}$$

The probabilities on the right-hand side are the parameters of the unconstrained LCM. The probability  $\pi_{x(i),y(i)}$  can be written in terms of the parameters of the LCM. First,

$$\pi_{x(i),y(i)} = \sum_{u=x}^m \sum_{v=y}^m P(X_i = u, X_i = v). \tag{14}$$

Under the LCM (Equation 13), the two scores on item  $i$  are locally independent, so that Equation 14 is equal to

$$\pi_{x(i),y(i)} = \sum_{u=x}^m \sum_{v=y}^m \sum_{k=1}^K P(\xi = k) P(X_i = u | \xi = k) P(X_i = v | \xi = k). \tag{15}$$

Second, inserting Equation 15 into Equation 3 gives

$$\begin{aligned} \text{LCRC} = & \frac{\sum_{i \neq j} \sum_x \sum_y [\pi_{x(i),y(j)} - \pi_{x(i)}\pi_{y(j)}]}{\sigma_X^2} \\ & + \frac{\sum_i \sum_x \sum_y \left[ \sum_{u=x}^m \sum_{v=y}^m \sum_{k=1}^K P(\xi = k) P(X_i = u | \xi = k) P(X_i = v | \xi = k) - \pi_{x(i)}\pi_{y(i)} \right]}{\sigma_X^2}. \end{aligned}$$

To estimate the LCRC, the researcher has to choose the number of latent classes,  $K$ , to obtain a good fit of the model to the data. The choice of the optimal  $K$  thus has to be based on statistical criteria. Because for medium and large numbers of variables, traditional goodness-of-fit statistics such as the likelihood ratio statistic  $G^2$  or Pearson's chi-square statistic  $X^2$  fail to provide trustworthy fit results (e.g., Koehler & Larntz, 1980), one usually resorts to relative fit measures, such as the information criteria AIC (Bozdogan, 1987) and BIC (Schwarz, 1978). Recently, Kang and Cohen (2007); Kang, Cohen, and Sung (2009); and Li, Cohen, Kim, and Cho (2009) evaluated several relative fit measures including AIC and BIC for choosing the correct item response theory model. The choice is made as follows. One selects a set of models and computes an information criterion for each model. The model yielding the lowest information criterion value is retained. Two of the three studies suggested using either AIC or BIC for choosing the best item response theory model, and the other study suggested using BIC.

The procedure for choosing an LCM using information criteria is similar. One starts with estimating the LCM for one class and computes the information criterion, then for two classes, three classes, and so on. As the number of classes increases, the information criterion value decreases until its minimum value, and then increases again. One stops estimating new LCMs when the information criterion value starts increasing again. The LCM yielding the lowest information criterion value is retained and used for computing the LCRC.

In the context of latent class analysis, another information criterion often used is AIC3 (Bozdogan, 1992). AIC3 has not been discussed in psychological measurement. Let  $LL$  be the estimated log likelihood of the LCM, and  $P$  the number of nonredundant parameters; that is,  $P = (K - 1) + JK(m - 1)$ . Then

$$\text{AIC3} = -2 \times LL + 3 \times P.$$

A series of simulation studies for various LCMs (Andrews & Currim, 2003; Dias, 2006; Lukočienė & Vermunt, 2010) showed that AIC tends to overestimate  $K$ , BIC tends to underestimate  $K$ , and AIC3 performed reasonably well. In this study, AIC3 was used to determine  $K$ .

## Comparing Five Methods to Estimate Reliability

A simulation study was used to compare accuracy and bias relative to the reliability, for alpha, lambda-2, MS, and LCRC, and one additional method, which is the split-half reliability coefficient based on random splits (SH-RS; Lord & Novick, 1968, p. 135). Method SH-RS does not fit into the present framework, but it was included because it is another single-administration method sometimes used by test constructors. SH-RS is computed by first splitting a test at random into two halves of equal length, computing the correlation between the total scores on the two half tests, and then using the Spearman-Brown prophecy formula to estimate the reliability of the total score on the whole test. If the test halves are parallel (Lord & Novick, 1968, p. 135), the outcome estimates the reliability; otherwise, underestimates or overestimates may be obtained. Revelle's beta provides the lowest split-half reliability, severely underestimating reliability, and Guttman's lambda-4 provides the highest split-half reliability. Guttman's lambda-4 often overestimates the reliability because of capitalization on the chance characteristics of samples (Thompson, Green, & Yang, 2010). The split-half reliability is available from most major statistical packages, for example, for the first and the second half of the items but, to the authors' knowledge, not based on random splits.

The five methods were compared under several conditions typical for test data. The following questions were investigated: (a) Is the bias of coefficients alpha and lambda-2 relative to  $\rho_{XX'}$  small enough to advocate these coefficient for practical use? (b) Is method MS unbiased when items are polytomous, given that the double monotonicity model does not hold? (c) Does method LCRC have smaller bias and greater accuracy than method MS?

## Method

The bias and the accuracy of the five reliability estimation methods were investigated using simulated data sets consisting of either dichotomous or polytomous item scores. Let  $\theta = (\theta_1, \dots, \theta_Q)$  be the  $Q$ -dimensional latent variable vector, with a  $Q$ -variate standard normal distribution. Let  $\psi_{jq}$  be the discrimination parameter of item  $j$  for latent variable  $q$ , and let  $\delta_{jx}$  be the location parameter for category  $x$  ( $x = 1, 2, \dots, m$ ) of item  $j$ . The multidimensional graded response model (De Ayala, 1994) is defined as

$$P(X_j \geq x|\boldsymbol{\theta}) = \frac{\exp \left[ \sum_{q=1}^Q \psi_{jq}(\theta_q - \delta_{jx}) \right]}{1 + \exp \left[ \sum_{q=1}^Q \psi_{jq}(\theta_q - \delta_{jx}) \right]}. \tag{16}$$

This model and the  $Q$ -variate standard normal  $\boldsymbol{\theta}$  were used to generate item scores and to compute the population reliability  $\rho_{XX'}$ .

Item scores for a sample of  $N$  simulees were generated as follows.  $N$  latent-variable vectors,  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ , were randomly drawn from the  $\boldsymbol{\theta}$  distribution. For each simulee (simulees are indexed  $n$ ) and each item, the  $m$  cumulative response probabilities were computed using Equation 16, and then the item score was randomly drawn from a multinomial distribution using the  $m$  cumulative response probabilities. Reliability  $\rho_{XX'}$  was closely approximated using a sample of 1 million simulees. For each latent-variable vector, the item scores were generated and total score  $X$  was computed. For each  $\boldsymbol{\theta}_n$ , the true score was computed as

$$T|\boldsymbol{\theta}_n = \sum_{j=1}^J E(X_j|\boldsymbol{\theta}_n) = \sum_{j=1}^J \sum_{x=1}^m P(X_j \geq x|\boldsymbol{\theta}_n),$$

where  $P(X_j \geq x|\boldsymbol{\theta}_n)$  is determined by Equation 16. Finally,  $\rho_{XX'}$  was computed using Equation 1.

The following design factors were varied:

*Reliability method (S).* The methods alpha, lambda-2, MS, LCRC, and SH-RS were studied.

*Test length (J).* The numbers of items were 6 (short test) and 18 (long test).

*Item format (m + 1).*  $J$  item scores were either dichotomous ( $m + 1 = 2$ ) or polytomous ( $m + 1 = 5$ ).

*Discrimination parameter ( $\psi$ ).* Discrimination parameters either differed across items (in which case they were inconsistent with the double monotonicity model) or they were equal (then they were consistent).

*Dimensionality (Q).* Unidimensional ( $Q = 1$ ) and two-dimensional ( $Q = 2$ ) latent variables were studied.  $Q = 1$  is consistent and  $Q = 2$  is inconsistent with the double monotonicity model.

*Sample size (N).* Samples were small ( $N = 200$ ) or large ( $N = 1,000$ ).

Reliability coefficient is a within-group factor, and the other factors are between-group factors. The standard case is defined as the comparison of bias and accuracy of the five reliability estimates for a short dichotomous-item test, generated for a large sample under Equation 16 with equal discrimination parameters and unidimensional  $\boldsymbol{\theta}$ . The standard case was compared to special cases, in which one of the conditions was varied relative to the standard case. Each comparison was replicated 1,000 times. The factors test length, item format, discrimination parameter, and dimensionality affect the choice of the item parameters of the multidimensional graded response model. Table 3 shows the item parameters for the standard case and the special cases of polytomous items, discrimination parameters differing across items, and two-dimensional latent variables. For long tests, the item-parameter values for Items 7 to 12 and 13 to 18 are equal to those for Items 1 to 6.

The dependent variables were bias and accuracy. Let  $S_b$  denote a reliability estimate in replication  $b$  ( $b = 1, \dots, B$ ), then the bias over  $B$  replications was computed as

**Table 3.** Item Parameters of Multidimensional Graded Response Model

Item	Standard		Polytomous				
	$\psi_j$	$\delta_j$	$\psi_j$	$\delta_{j1}$	$\delta_{j2}$	$\delta_{j3}$	$\delta_{j4}$
1	1	-2.5	1	-4	-3	-2	-1
2	1	-1.5	1	-3	-2	-1	0
3	1	-0.5	1	-2	-1	0	1
4	1	0.5	1	-1	0	1	2
5	1	1.5	1	0	1	2	3
6	1	2.5	1	1	2	3	4

  

Item	Unequal $\psi$		2 Dimensions		
	$\psi_j$	$\delta_j$	$\psi_{j1}$	$\psi_{j2}$	$\delta_j$
1	0.5	-2.5	1	0	-2.5
2	2	-1.5	1	0	-1.5
3	0.5	-0.5	1	0	-0.5
4	2	0.5	0	1	0.5
5	0.5	1.5	0	1	1.5
6	2	2.5	0	1	2.5

Note: Unequal  $\psi$  = discrimination parameters differ across items.

$$\text{bias} = \frac{1}{B} \sum_{b=1}^B (S_b - \rho_{XX'}). \quad (17)$$

Absolute bias was interpreted as follows:  $|\text{bias}| < .001$  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. For assessing accuracy, the mean absolute error (MAE) was used, which was defined as

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

MAE provides information on the error one can expect for a single data set. The MAE was interpreted as follows:  $\text{MAE} < .002$  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 simulations were done in R (R Development Core Team, 2006). The computer code is available on request from the first author. Coefficients alpha, lambda-2, MS, and LCRC are available from the R-package *mokken* (version 2.5 and higher; Van der Ark, 2007).

## Results

The number of latent classes,  $K$ , required for computing each of the 6,000 LCRCs ranged from 2 to 5, with a modal value of  $K = 3$ . Table 4 shows  $\rho_{XX'}$  values, and the bias and the MAE of the alpha, lambda-2, MS, LCRC, and SH-RS estimates. Alpha and the SH-RS had the largest bias, which ranged from small (long-test condition for both alpha and SH-RS, and polytomous-items condition for SH-RS) to large (two-dimensional data). Estimates lambda-2 and MS were almost unbiased for data based on equal discrimination parameters (i.e., consistent with double monotonicity) for both dichotomous and polytomous items. However, bias was large when data were

**Table 4.** Bias and MAE of Five Reliability Estimation Methods

Condition	$\rho_{XX'}$	Bias				
		Alpha	Lambda-2	MS	LCRC	SH-RS
Standard	.464	-.018	-.001	.004	-.010	-.011
Polytomous	.765	-.015	-.001	-.001	-.009	-.009
Unequal $\psi$	.424	-.045	-.030	-.027	-.012	-.036
2 dimensions	.315	-.080	-.049	-.031	-.020	-.083
Long test	.722	-.009	-.003	-.000	-.004	-.005
Small $N$	.464	-.021	-.004	.002	-.006	-.015

  

Condition	MAE				
	Alpha	Lambda-2	MS	LCRC	SH-RS
Standard	.025	.022	.024	.022	.029
Polytomous	.015	.011	.009	.011	.015
Unequal $\psi$	.047	.034	.034	.024	.044
2 dimensions	.080	.051	.040	.045	.092
Long test	.012	.010	.010	.011	.015
Small $N$	.046	.042	.048	.042	.059

Note: Unequal  $\psi$  = discrimination parameters differ across items.

not unidimensional or discrimination parameters were unequal (i.e., inconsistent with double monotonicity). Only the LCRC method had no considerable or large bias in any of the conditions. For all conditions, the bias was largest for 2-dimensional data and for data generated under a graded response model with unequal discrimination parameters. Bias was smallest for the condition with a large number of items. Sample size and item format did not affect bias. Furthermore, also for the MS and LCRC methods, the bias was predominantly negative.

Differences in accuracy due to condition were greater than differences due to reliability estimation method. Reliability was estimated most accurately for polytomous items and long tests (small MAE). Reliability was estimated least accurately for two-dimensional data and small sample size (MAE had considerable or high values). Alpha and SH-RS were less accurate than lambda-2, MS, and LCRC. For unequal discrimination parameters, the LCRC method was more accurate than the other methods.

## Discussion

The alpha, lambda-2, and MS methods were cast in the same theoretical framework. A new reliability method, LCRC, was proposed in the context of this theoretical framework. Theoretically, the LCRC method is superior to the other methods because the terms in the unobserved numerator in Equation 3 are estimated with fewer restrictions. Hence, restrictive conditions such as essential  $\tau$ -equivalence and double monotonicity are not prohibitive in finding estimates with little bias.

The simulation study showed that for all conditions, the alpha and SH-RS methods have potentially large bias and MAE. The authors recommend not using these methods when better alternatives are available. If the double monotonicity model does not hold (i.e., discrimination parameters differ across items, or the data are multidimensional), LCRC is less biased relative to  $\rho_{XX'}$  than the other methods, otherwise lambda-2 and MS are less biased. For accuracy, the picture is not as clear as for bias. LCRC is most accurate for varying discrimination parameters, but MS is slightly more accurate for multidimensional data. If it is unknown whether data are

unidimensional or the items have equal discrimination parameters, LCRC is a good choice; otherwise lambda-2 and MS are good choices.

The information measure AIC3 was used in the present study for determining the number of latent classes needed for computing the LCRC, but more research has to be done to find the best information measure. A drawback of all information criteria is that they are *relative* fit measures. The LCM yielding the lowest information criterion value fits best relative to other LCMs but still may have a poor *absolute* fit. Absolute fit assessment may be improved by inspecting the bivariate or trivariate residuals, but a methodology for dealing with these residuals is currently not available.

### Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

### Funding

The author(s) received no financial support for the research and/or authorship of this article.

### References

- 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. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137-143.
- Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*, 249-267.
- Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.
- Bozdogan, H. (1992). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and classification: Concepts, methods and applications* (pp. 40-54). New York, NY: Springer.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, *78*, 98-104.
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.
- De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*, 155-170.
- 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.), *Studies in classification, data analysis, and knowledge organization* (pp. 91-199). Berlin, Germany: Springer.
- Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121-135.
- 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.
- Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*, 331-358.
- Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*, 499-518.

- Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*, 336-344.
- 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.), *Measurement and prediction* (pp. 362-412). Princeton, NJ: Princeton University Press.
- Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353-373.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lukočienė, 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.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: Mouton; Berlin, Germany: De Gruyter.
- Molenaar, I. W. (1997). Nonparametric models for polytomous items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369-380). New York, NY: Springer.
- Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden*, *9*(28), 115-126. Retrieved from <http://arno.uvt.nl/show.cgi?fd=81058>
- Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. A program for Mokken scale analysis for polytomous items*. Groningen, Netherlands: iec ProGAMMA.
- R Development Core Team. (2006). R: A language and environment for statistical computing [computer programming language]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173-184.
- Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*, 195-212.
- Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*, 57-74.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.
- Sedere, M. U., & Feldt, L. S. (1977). The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's lambda-2. *Journal of Educational Measurement*, *14*, 53-62.
- Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory* (Unpublished doctoral dissertation). Amsterdam, Netherlands: Vrije Universiteit.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.
- Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test score in nonparametric item response theory. *Psychometrika*, *52*, 79-97.
- 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., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*, 575-579.
- Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the maximal split-half coefficient to estimate reliability. *Educational and Psychological Measurement*, *70*, 232-251.
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1-19.
- Van der Ark, L. A. (2010). Computation of the Molenaar Sijtsma statistic. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 775-784). Berlin, Germany: Springer.
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's  $\alpha$ , Revelle's  $\beta$ , and McDonald's  $\omega$ : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123-133.