DOCUMENT RESUME

ED 389 754 TM 024 387

AUTHOR Meijer, Rob R.; And Others
TITLE Reliability Estimation for Single Dichotomous Items. Research Report 94-5.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE Nov 94
NOTE 32p.
AVAILABLE FROM Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); Foreign Countries; *Item Response Theory; Monte Carlo Methods; Nonparametric Statistics; *Research Methodology; Sampling; Statistical Bias; Test Construction; Test Items; *Test Reliability
IDENTIFIERS *Dichotomous Variables; Person Fit Measures

ABSTRACT

Three methods for the estimation of the reliability of single dichotomous items are discussed. All methods are based on the assumptions of nondecreasing and nonintersecting item response functions and the Mokken model of double monotonicity. Based on analytical and Monte Carlo studies, it is concluded that one method is superior to the other two because it has a smaller bias and a smaller sampling variance. Furthermore, this method shows some robustness under violation of the condition of nonintersecting item response functions. Item reliability is of special interest for Mokken's nonparametric item response theory and is useful for the evaluation of item quality in nonparametric test construction research. It is also of interest for nonparametric person fit analysis. (Contains 1 figure, 2 tables, and 33 references.) (Author/SLD)

Reliability Estimation for Single Dichotomous Items

Research Report 94-5

Rob R. Meijer

Klaas Sijtsma

Ivo W. Molenaar

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Reliability Estimation for Single Dichotomous Items based on Mokken's IRT Model

Rob R. Meijer
Klaas Sijtsma
Ivo W. Molenaar

Reliability Estimation for Single Dichotomous Items based on Mokken's IRT Model, Rob R. Meijer, Klaas Sijtsma, Ivo W. Molenaar - Enschede: University of Twente, Faculty of Educational Science and Technology, November, 1994, - 26 pages.

Abstract

Three methods for the estimation of the reliability of single dichotomous items are discussed. All methods are based on the assumptions of nondecreasing and nonintersecting item response functions. Based on analytical and Monte Carlo studies, it is concluded that one method is superior over the other two, because it has a smaller bias and a smaller sampling variance. Furthermore, this method shows some robustness under violation of the condition of nonintersecting item response functions. Item reliability is of special interest for Mokken's nonparametric item response theory, and is useful for the evaluation of item quality in nonparametric test construction research. It is also of interest for nonparametric person fit analysis.

Key words: item reliability, item response theory, Mokken model, nonparametric item response models, test construction.

Introduction

For the practical use of tests total scores are more important than scores on individual items. In test construction, however, the quality of items must be assessed to select the appropriate items that, taken together, constitute a useful test. For example, in classical test theory (CTT; Lord & Novick, 1968) item statistics like the p-value and the corrected item-total correlation are used for this purpose. In logistic item response theory (IRT; e.g., Lord, 1980) items can be evaluated on the basis of their difficulty, discrimination power, and pseudo-chance level. Moreover, the item information function (Lord, 1980, p. 72) can be used to assess measurement accuracy of the individual item. The nonparametric Mokken (1971, 1994; Mokken & Lewis, 1982) approach to IRT uses the p-value and an item scalability coefficient.

Because the Mokken approach provides the theoretical framework for this study, we will further concentrate on its relevant assumptions and definitions. We will argue that in the Mokken IRT approach the reliability of an item can serve as a nonparametric counterpart of the discrimination power from logistic IRT and the corrected item-total correlation from CTT [refer to Lord (1980, p. 33) for a comparison of these latter two item statistics].

The purpose of this paper is to apply three relatively simple methods, used earlier for the estimation of total score reliability in the nonparametric Mokken IRT framework (Mokken, 1971, pp. 142 - 147; Sijtsma & Molenaar, 1987), to the estimation of single item reliability. The asymptotic bias and the finite sample bias of these methods will be investigated. Furthermore, their standard deviation

will be studied, whereas results pertaining to skewness and kurtosis will be briefly summarized.

The Nonparametric Mokken Approach and Item Reliability

Nonparametric IRT models are important because of their potential to order persons and items. Cliff and Donoghue (1992) provide arguments that favor ordinal rather than interval measurement in psychological and educational testing. Mokken (1971, pp. 115 - 169; 1994; Mokken & Lewis, 1982) proposed two nonparametric IRT models for the analysis of binary item scores. The first is the model of monotone homogeneity (MH) which is defined by three assumptions: Unidimensionality, local stochastic independence, and nondecreasingness of the item response functions (IRFs). An important property of the MH model is that the latent trait score is stochastically ordered by the number-correct score on k items (Grayson, 1988; Huynh, 1994). Similar models were studied by Holland (1981), Rosenbaum (1984), Stout (1990), Ellis and van den Wollenberg (1993), and Junker (1993); other ordinal models by Schulman and Haden (1975) and Cliff (1979).

The second model is the model of double monotonicity (DM). This model rests on the same three assumptions as the MH model, plus the fourth assumption that the IRFs do not intersect. Thus, the DM model not only allows persons to be ordered, but also allows an ordering of items that is identical, except for possible ties, for all persons taking the test. Similar models were discussed by Rosenbaum (1987), Croon (1991), Sijtsma and Meijer (1992), and Sijtsma and Junker (1994).

It may be noted that the Rasch (1960) model is based on the three assumptions from the MH model, plus the fourth assumption of minimal sufficiency of the number-correct scores of persons and items for the estimation of the latent person and item parameters, respectively (Fischer, 1974, pp. 193 - 203). Not only are the IRFs from the Rasch model strictly increasing and nonintersecting but they are also parallel. Levine (1970) has discussed conditions from which it can be derived that, in general, DM IRFs can not be transformed into Rasch IRFs. For example, the DM model allows IRFs with asymptotes that are unequal to 0 or 1 whereas the Rasch model excludes such IRFs. Disregarding the trivial case of constant IRFs, theoretically, the DM model includes the Rasch model as a special case. In practice, however, differences will become apparent in particular for small numbers of items. For larger numbers the DM model still allows relatively easy items to have pseudo-chance levels larger than 0 and relatively difficult items to have upper asymptotes smaller than 1. This is not at all unrealistic because easy items may also be relatively easy for low-ability examinees, even if there is no guessing, while difficult items need not be trivial for high-ability examinees. Meijer, Sijtsma, and Smid (1990) provide a theoretical and a practical comparison of the DM and the Rasch model.

Note that with their nonparametric definition of the IRFs, the MH and the DM models do not assume particular distributions for latent model parameters. In other words, characteristics of the models hold irrespective of such distributions. As a result of a nonparametric definition, latent item parameters from parametric models, such as the item difficulty and the discrimination power, can not be numerically estimated. In the nonparametric approach by Mokken (1971; Mokken & Lewis, 1982), the latent item difficulty is replaced by the proportion of correct responses given on an item (Mokken, 1971, p. 124). Furthermore, Mokken (1971,

p. 151; Mokken & Lewis, 1982) proposed an item coefficient that expresses the scalability of a particular item with respect to the scale of the other items. Mokken, Lewis, and Sijtsma (1986) noted that this coefficient is also related to the slope of an IRF. In addition, Donoghue and Cliff (1991) noted that the Mokken approach does not provide much specific information at the item level. An item statistic that is more directly related to the discrimination power could be useful in item selection. Such a statistic can also play a useful role in nonparametric person fit analysis (e.g., Meijer, Molenaar, & Sijtsma, 1993; Tatsuoka & Tatsuoka, 1983; van der Flier, 1982). In this study, item reliability is proposed as an appropriate replacement for discrimination power [also refer to Meredith (1965) for a similar proposal] in a nonparametric IRT context. This can be explained as follows.

The reliability of an item expresses the degree to which observed item scores can be repeated independently under similar conditions. Discrimination power (denoted by α) as defined in logistic IRT (Lord, 1980) has a similar interpretation. Let Θ be the latent person parameter with probability density $f(\Theta)$. Furthermore, let item g have a latent difficulty parameter δ_g and a latent discrimination parameter α_g. Keeping $f(\Theta)$ and δ_g fixed, an increase in α_g corresponds to a higher degree of repeatability of observed scores on item g. In the limit ($\alpha_g \rightarrow \infty$), response performance is in accordance with the deterministic Guttman (1950) model: this means perfect repeatability and thus perfect item reliability. For response behavior following a logistic IRT model, an increase in α_g yields lower probabilities of a correct response to the left of δ_g and higher probabilities to the right of it. Consequently, for each subject with $\Theta \neq \delta_g$ his/her dominant item response (which is incorrect for $\Theta < \delta_g$ and correct for $\Theta > \delta_g$) can be predicted with higher probability. Note, that for $\Theta = \delta_g$ the

success probability is a constant irrespective of α_g. In other words, holding everything else constant, an increase in α_g corresponds to a higher degree of repeatability of item scores.

Definition and Estimation of Item Reliability

Because the theoretical basis for the definition and the estimation of item reliability was given by Mokken (1971, pp. 142 - 147), and Sijtsma and Molenaar (1987), we will only provide results here. Let π_g be the population proportion of persons giving a correct response on the dichotomous item g, and π_{gg} the population proportion giving a correct response on two locally independent replications of item g. As a tool for estimating the reliability of a test score, Mokken (1971 p. 143) defines the reliability of the dichotomous item score X_g as

$$\rho(X_g) = \frac{\pi_{gg} - \pi_g^2}{\pi_g(1-\pi_g)} = 1 - \frac{\pi_g - \pi_{gg}}{\pi_g(1-\pi_g)} . \quad (1)$$

Reliability equal to 0 is obtained if $\pi_{gg} = \pi_g^2$ (statistical independence between replications of item g); reliability equal to 1 if $\pi_{gg} = \pi_g$.

The proportion π_g can be estimated unbiasedly (Mokken, 1971, p. 126) but because locally independent replications of items usually are absent, a direct

estimate of π_{gg} is not available. Therefore, Mokken (1971, p. 143) proposed two methods using parameters for which sample estimators are available to approximate π_{gg}. Sijtsma and Molenaar (1987) proposed a third method. All three methods are based on extrapolation or interpolation using items adjacent to item g in the ordering of items from difficult to easy. The rationale is the following.

Assume that the k items from the test are ordered according to increasing π_g and that item indices are in accordance with this ordering. Let the IRFs denoted by $\pi_g(\Theta)$ (g = 1, ..., k) of all k items be nonintersecting: for items g-1, g, and g+1 this means that

$$\pi_{g-1}(\theta) \leq \pi_g(\theta) \leq \pi_{g+1}(\theta), \textit{ for all } \theta \ . \tag{2}$$

Based on the idea that the IRFs of the neighbor items in the item ordering are more similar to $\pi_g(\Theta)$ than the other IRFs, all three methods use either $\pi_{g-1}(\Theta)$ or $\pi_{g+1}(\Theta)$, or both as a predictor of a real replication of item g. Note that π_{gg} equals

$$\pi_{gg} = \int_{\theta} \pi_g(\theta)\pi_g(\theta)dF(\theta) \ . \tag{3}$$

Before integrating with $dF(\Theta)$, one of the probabilities $\pi_g(\Theta)$ is replaced by a linear approximation using one or two of its neighbors, $\pi_{g-1}(\Theta)$ or $\pi_{g+1}(\Theta)$, or both: $\tilde{\pi}_g(\Theta) = a + b\pi_{g-1}(\Theta) + c\pi_{g+1}(\Theta)$. Each method is defined by the choice of a, b, and c. Substitution of $\tilde{\pi}_g(\Theta)$ in (3) and integration yield

$$\tilde{\pi}_{gg} = a\pi_g + b\pi_{g-1,g} + c\pi_{g,g+1} . \quad (4)$$

In (4), $\pi_{g-1,g}$ is the population proportion of persons that have correct responses on both items g-1 and g. A similar definition applies to $\pi_{g,g+1}$. Mokken's (1971, p. 147) method 1 uses extrapolation with π_g, π_{g-1}, and $\pi_{g-1,g}$, or π_g, π_{g+1}, and $\pi_{g,g+1}$. Sijtsma and Molenaar (1987) provided two alternative approximations to π_{gg}. Because each of these four approximations is asymptotically biased (Molenaar & Sijtsma, 1984), Sijtsma and Molenaar's (1987) method used the unweighted mean of these four approximations which has only small bias.

Mokken's method 2 uses both neighbors of item g to approximate π_{gg} by interpolation (Mokken, 1971, p.147). For the two extreme items extrapolation (method 1) is used. Refer to Sijtsma and Molenaar (1987) for further details. Note that these earlier publications only give results pertaining to sample bias and variance of total score reliability estimation for each of the three reliability methods: ρ_1, ρ_2, and ρ_{MS}.

All approximations to π_{gg} are functions of the bivariate proportions and the distance between item difficulties. If a bivariate proportion is smaller or a distance is larger than expected if the items had been replications this may bias $\tilde{\pi}_{gg}$ and, consequently, the reliability estimate of item g. Figure 1 illustrates the effect of distance on the approximation of $\pi_g(\Theta)$ by means of one neighbor (method 1; left in Figure

Insert Figure 1 about here

1) or two neighbors (method 2; right in Figure 1). To illustrate estimation of π_{gg} using method 1 we need the extrapolation formula (Mokken, 1971, p. 147)

$$\tilde{\pi}_{gg} = \frac{\pi_{g,g+1}\pi_{g}}{\pi_{g+1}} \quad . \tag{5}$$

For method 1, the striped curve denoted $\tilde{\pi}_{g}(\Theta)$ in Figure 1 (left) gives the approximation to $\pi_g(\Theta)$ using $(\pi_g/\pi_{g+1})\pi_{g+1}(\Theta)$. [note that substitution of this product in (3) yields (5)]. Assume that Θ follows a normal distribution with its peak at the scale value for which $\pi_g(\Theta) = .5$. The approximation on the basis of $\pi_{g+1}(\Theta)$ overestimates $\pi_g(\Theta)$ to the left of the scale, but it underestimates $\pi_g(\Theta)$ to the right of it. As it is multiplied by the factor $\pi_g(\Theta)dF(\Theta)$, higher values of Θ tend to contribute most to the integral that yields the approximation to π_{gg} in (5). The underestimation thus tends to dominate the overestimation. A larger distance usually results in a worse approximation. If $\pi_{g+1}(\Theta)$ - $\pi_g(\Theta)$ increases while keeping $\pi_g(\Theta)$ fixed, the multiplication factor π_g/π_{g+1} in (5) decreases and the approximation to $\pi_g(\Theta)$ lies further to the left of $\pi_g(\Theta)$ and also further below it at the right side of the scale. Thus, it tends to underestimate $\pi_g(\Theta)$ more strongly if the distance is larger. The same line of reasoning leads to the conclusion that the approximation based on $\pi_{g-1}(\Theta)$ (formula not given here) tends to overestimate $\pi_g(\Theta)$ and, as a result, $\tilde{\pi}_{gg}$ more strongly overestimates π_{gg} if the curves $\pi_{g-1}(\Theta)$ and $\pi_g(\Theta)$ lie further apart.

For method 2 (Figure 1, right; formula not given here), the underestimation at the right of the scale obtains a larger weight than the overestimation at the left, and $\tilde{\pi}_{gg}$ according to method 2 tends to be an underestimate. Moving $\pi_{g-1}(\Theta)$ further to the right while keeping $\pi_g(\Theta)$ and $\pi_{g+1}(\Theta)$ fixed, thus increasing inequality of distances leads to a situation in which it is difficult to predict how

the bias of $\hat{\pi}_{gg}$ will be affected.

These examples lead to the conclusion that distance affects the degree to which $\hat{\pi}_{gg}$ is biased, and unequal distances of both neighbors to $\pi_g(\Theta)$ affects the bias differently than equal distances. Given the susceptibility of the item reliability methods to the quality of other items in the test, it will be investigated which of the three methods has the smallest bias.

An alternative approach would be the use of the m ($m > 2$) nearest neighbors to approximate π_{gg}. However, neighbors that are farther away are less similar (in the sense of replications) to item g than the two nearest neighbors. Thus, we would expect larger bias in estimating π_{gg} for $m > 2$. By using more information from the data, however, the sampling variance of the estimates might decrease compared with $m = 2$. An acceptable compromise between bias and accuracy would, probably, depend on several characteristics of test, items, and population. Also refer to Donoghue and Cliff (1991) and Cliff and Donoghue (1992) who use ordinal multiple regression for a related problem in ordinal true score theory. Rather than pursuing a more complex strategy, we will stay within the confines of the Mokken approach and investigate asymptotic and sampling characteristics of reliability estimators based on the simpler methods 1, 2, and MS. Only if none of these methods yields satisfactory results may a more complex strategy be rewarding.

An analytical derivation of the distribution properties of the three methods is not pursued because the ordering of the items according to their difficulty may well vary across random samples and different approximations to π_{gg} will be used. Therefore, conclusions will be based on simulation studies.

Asymptotic Bias in Item Reliability Estimation Methods

Method. As a first step, the bias of each of the three item reliability methods with respect to $\rho(X_g)$ in (1) was investigated using population fractions obtained via numerical integration across the ability distribution. This allowed to study the performance of the three methods in the ideal case of very large samples. Throughout this study sets of 7 items were used. Such a small set was large enough because (1) the focus of attention was on the individual item; (2) distance between items could be manipulated equally well in small and large sets; (3) differences between extremely located items and items in between could be studied independently of test length; and (4) with usually smaller distances between adjacent items in longer tests, results for smaller tests were expected to be conservative. Furthermore, logistic IRFs were used. Note, in particular, that although our theoretical framework is nonparametric IRT, parametrically defined IRFs and parameter distributions are necessary to simulate 0's and 1's.

Given 7 two-parameter logistic IRFs and a standard normal distribution of Θ, numerical integration (IMSL routine DCADRE, 1982) was used to obtain the population proportions π_g (g = 1, ..., 7), π_{gg} (g = 1, ..., 7) and π_{gh} (g, h = 1, ..., 7; $g \neq h$). Using π_g and π_{gg}, the item reliability $\rho(X_g)$ was calculated. To calculate item reliability with approximation methods 1, 2 and MS, the proportions π_g and π_{gh} were used: the results are denoted by ρ_1, ρ_2, and ρ_{MS}, respectively. The difference between each of these parameters and $\rho(X_g)$ equals the bias of a specific method with respect to the reliability (1) for item g (g = 1, ..., 7).

A completely crossed 4 x 2 x 3 design was used. The first factor was average discrimination power α_M (subscript M denotes mean), with four levels: α_M = .5, 1, 2, and 5. In combination with a standard normal distribution of Θ these values cover the complete range from very weak to very strong discrimination (Meijer et al., 1993). The second factor was spread of the αs within one test, with two levels: zero spread (all 7 αs equal) and positive spread (αs unequal). Zero spread corresponds to nonintersection of two-parameter logistic IRFs. For example, for α_M = 1 we have α_g = 1 (g = 1, ..., 7). Positive spread corresponds to intersection of IRFs, and thus provides a violation of a condition underlying estimation of item reliability. For example, for α_M = 1 we have α = (1.3, 1, 1, .7, 1, 1.3, .7). This more realistic condition allows us to investigate the robustness of the estimation methods. The third factor was distance between item locations. A distinction was made between sets of equally spaced items and sets of unequally spaced items. Three levels were distinguished. On two levels, item locations (δs) were equidistant with median equal to zero and distance [d(δ)] equal to either .1 or .5, respectively. These levels are denoted ES (Equidistant, Small distance) and EL (Equidistant, Large distance), respectively. On the third level, d(δ) varied more realistically within one item set. In particular, δ = (-.4, -.3, -.2, 0, .2, .8, 1.6) for all design cells on this level. The third level is denoted UD (Unequal Distance).

Results. Table 1 shows a summary of the asymptotic bias results for the complete design. For Nonintersecting IRFs (left half of Table 1) the

Insert Table 1 about here

general conclusion for method MS is that the reliability is almost unbiased for most items (results denoted by #nobi, number of items having "no bias"). Out of 84 reliabilities (12 cells), 70 have a bias smaller than |.01|, and 75 have a bias smaller than |.03|. The largest bias (denoted min, for α_M = 5 and UD) is -.06. The results for #nobi, min and max are almost always better for method MS than for methods 1 and 2. These latter methods often yield unacceptably large biases, for example, larger than |.10|. Method 1 often has a large bias for most of the 7 items in the test. Method 2 mostly yields large biases for the two extreme items (for which, in fact, method 1 is used) and sometimes also for the items in between.

For intersecting IRFs (right half of Table 1) asymptotic bias is larger for all three methods. For method MS, 25 of the 84 reliabilities have a bias smaller than |.01|, and 53 have a bias smaller than |.03|. The largest bias is -.07 (min for α_M = 5 and UD). As for Nonintersecting IRFs, the results for Intersecting IRFs are almost always better for method MS than for the other two methods. With a few exceptions (not all individual values are shown in Table 1), the bias of method MS for individual item reliabilities is acceptable.

For method MS, the grand mean of the bias is .001. Main effects and interaction effects are mostly very close to 0 (< |.01|), with one exception for α_M = 5 and EL (first-order interaction is -.03).

It can be concluded for method MS that: (1) bias is smaller than for methods 1 and 2; (2) bias is often negligible or practically acceptable; and (3) bias stays within reasonable limits even if IRFs intersect.

Finite Sample Estimation of Item Reliability

Method. A Monte Carlo study was conducted to assess the sampling characteristics of the three approximations to item reliability for realistic sample sizes. Despite the larger asymptotic biases for method 1 and method 2 (Table 1), they were also subjected to the Monte Carlo investigation because: (1) not only bias is important but also sampling variance; (2) it could well happen that a method with larger asymptotic bias has smaller finite sample bias, given e.g. the additional problem of different neighbors mentioned above; and (3) methods 1 and 2 are simpler than method MS and might thus be recommended if the bias of method MS is only slightly larger.

Data matrices containing n(persons) x k(items) binary item scores were generated (for the simulation procedure see Sijtsma & Molenaar, 1987) using two-parameter logistic IRFs and a standard normal distribution of Θ. The design from the asymptotic bias study was extended by adding sample size as a fourth factor with three levels: n = 100, 300, and 900. The sample size n = 100 can be considered to be typical of ad hoc test construction that is part of a larger research project, n = 300 of test construction research as performed in a non-commercial environment (e.g., universities, where the means to collect data from larger samples are limited), and n = 900 (or more) of large scale test construction on a more commercial basis.

Thus a completely crossed 4 x 2 x 3 x 3 design was used. The number of replications in each cell of the design was 200. For each replication, the estimated π_g and π_{gh} were used (in the order found from that matrix) for

estimation of ρ by methods 1, 2, and MS.

Results. Method MS has almost always a smaller finite sample bias than methods 1 and 2. In addition, for practical purposes the bias of method MS can be ignored. Furthermore, the standard deviation of method MS is almost always smaller than that of methods 1 and 2 (not tabulated here). Because of these results only the Monte Carlo results for method MS are discussed.

In Table 2 (results for bias and standard deviation for n = 300), it can be seen that method

Insert Table 2 about here

MS is almost unbiased. For the widely spaced items (Table 2, EL) that have Nonintersecting IRFs (Table 2, first half) the bias is, except for $\alpha_M = 5$, somewhat larger for the extreme items. For $\alpha_M = 5$, the bias is larger for the items in between. For unequally spaced items (Table 2, UD), bias is negligible save a few exceptions if $\alpha_M = 2$ and $\alpha_M = 5$. For n = 100 (not tabulated), bias is in general somewhat higher, especially for $\alpha = .5$ and $\alpha = 1$. For n = 900 (not tabulated), bias results are highly comparable to the results obtained for n = 300.

For n = 300 and Nonintersecting IRFs (Table 2, first half), the standard deviation for almost all items is approximately .05. Only the standard deviation for the extremely easy and difficult items from widely spaced sets of items (Table 2, EL) sometimes is somewhat larger. For small sample size (n = 100; not tabulated), the standard deviation of method MS across samples is rather large (between .7 and .13 for the extreme items and between .04 and .09 for the items in between). For large sample size (n = 900; not tabulated), the standard

deviation for almost all items is approximately .025. In general, for n = 100 the standard deviation is approximately $\sqrt{3}$ times as large as for n = 300, and for n = 900 it is approximately $\sqrt{3}$ times as small as for n = 300.

The results for the third and fourth moments (not tabulated) are briefly summarized. For $\alpha_M = 1$ and $\alpha_M = 2$ the distribution of estimator MS is rather symmetrical around its mean (all sample sizes; skewness between -.4 and .4). For $\alpha_M = .5$, for some items the distribution is positively skewed. For $\alpha_M = 5$ (all sample sizes), the distribution is negatively skewed for some items and positively skewed for others. The peakedness of the distribution is more or less comparable with the normal distribution for all discrimination levels (in general, the kurtosis is approximately 3).

If IRFs intersect (Table 2, second half), the bias of method MS is generally larger than if IRFs do not intersect (Table 2, first half). The pattern of biases across items within a test is rather inconsistent. For a few items bias ranges from -.08 to .05. However, for the majority of the items bias is much smaller. Compared with nonintersection of IRFs (Table 2, first half), standard deviation results are approximately the same if IRFs intersect (Table 2, second half). The same conclusion holds for skewness and kurtosis results.

The grand mean of the bias is -.001 for the results pertaining to n = 300. The vast majority of main and interaction effects is smaller than |.01|. A few exceptions occur for some first, second, and third order interactions (effects smaller than |.02| in most cases; never larger than |.03|) for $\alpha_M = 5$. ANOVA results for the standard deviation show no interesting effects.

Discussion

This study has introduced and compared three methods (method 1, method 2, and method MS) for the estimation of the item reliability that are based on the Mokken model of double monotonicity. Method MS was unbiased for almost all items with the exception of a small and probably unimportant bias for the extreme items if item difficulties are widely spaced. In addition, in all cases studied, method MS had smaller bias than the other two reliability methods. Reduction of the bias of method MS for the extreme items seems problematic, because the use of, for example, the m ($m > 2$) nearest neighbor items rather than the two nearest neighbors would probably reduce the standard deviation but increase the bias, in particular for the two extreme items.

Method MS had the smallest standard deviation across random samples. For a sample size $n = 300$, its standard deviation ranged from .03 to .06. For small samples ($n = 100$), on the average the standard deviation was larger by a factor of approximately $\sqrt{3}$ and for larger samples ($n = 900$) it was smaller by the same factor. It may be concluded that for $n = 300$ and larger samples reliability estimates are accurate enough to allow the identification of unreliable items. For small samples ($n = 100$) accuracy may be too small, but it may be noted that such samples are generally considered to be too small for serious test construction and only allow tentative conclusions about the quality of a test and its items. Finally, other results indicated that the sampling distribution of method MS is approximately symmetrical in most situations that were considered here.

References

Cliff, N. (1979). Test theory without true scores? Psychometrika, 44, 373-393.

Cliff, N., & Donoghue, J. R. (1992). Ordinal test fidelity estimated by an item sampling model. Psychometrika, 57, 217-236.

Croon, M. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. British Journal of Mathematical and Statistical Psychology, 44, 315-331.

Donoghue, J. R., & Cliff, N. (1991). An investigation of ordinal true score test theory. Applied Psychological Measurement, 15, 335-351.

Ellis, J. L., & van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. Psychometrika, 58, 417-429.

Fischer, G. H. (1974). Einführung in die Theorie Psychologischer Tests. Bern: Huber.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. Journal of Cross-cultural Psychology, 13, 267-298.

Grayson, D. A. (1988). Two group classification in latent trait theory: Scores with monotone likelihood ratio. Psychometrika, 53, 383-392.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A Clausen (Eds.), Measurement and prediction (pp. 60-90). Princeton: Princeton University Press.

Holland, P. W. (1981). When are item response models consistent with observed data ? Psychometrika, 46, 79-92.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. Psychometrika, 59, 77-79.

IMSL Library (1982). Houston: IMSL.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. The Annals of Statistics, 21, 1359-1378.

Levine, M. V. (1970). Transformations that render curves parallel. Journal of Mathematical Psychology, 7, 410-443.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

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

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1993). Item, test, person and group characteristics and their influence on nonparametric appropriateness measurement. Applied Psychological Measurement (in press).

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. Applied Psychological Measurement, 14, 283-298.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. Psychometrika, 30, 419-440.

Mokken, R. J. (1971). A theory and procedure of scale analysis. New York, Berlin: de Gruyter.

Mokken, R. J. (1994). Nonparametric models for dichotomous items. In W.J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern test theory. New York: Springer Verlag (in press).

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 6, 417-430.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: A critical discussion'. Applied Psychological Measurement, 10, 279-285.

Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. Tijdschrift voor Onderwijsresearch, 9, 257-268.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Rosenbaum, P. R. (1987). Probability inequalities for latent scales. British Journal of Mathematical and Statistical Psychology, 40, 157-168.

Schulman, R. S., & Haden, R. L. (1975). A test theory model for ordinal measurements. Psychometrika, 40, 455-472.

Sijtsma, K., & Junker, B. W. (1994). A survey of theory and methods of invariant item ordering (submitted for publication).

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. Applied Psychological Measurement, 16, 149-157.

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

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, 7, 215-231.

Table 1

Asympt[illegible] Bias Results for Parameters ρ_1, ρ_2 and ρ_{MS} Relative to ρ [illegible]

		Uninteracting IRFs									Interacting IRFs								
		ES			EL			UC			ES			EL			UC		
		#n.bi	min	max	#n.bi	min	max	#n.bi	min	max	#n.bi	min	max	#n.bi	min	max	#n.bi	min	max
α[illegible]	ρ_1, ρ	[illegible]			[illegible]			[illegible]			[illegible]	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	ρ_2, ρ	[illegible]						[illegible]			[illegible]	.12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	ρ_{MS}, ρ	[illegible]			[illegible]			[illegible]			[illegible]	[illegible]	4	[illegible]	[illegible]	[illegible]	[illegible]		[illegible]
α[illegible]	ρ_1, ρ	[illegible]			[illegible]	4	4	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	4	[illegible]	[illegible]	[illegible]	4
	ρ_2, ρ	[illegible]			[illegible]	[illegible]	4	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4
	ρ_{MS}, ρ	[illegible]			[illegible]		[illegible]	[illegible]		[illegible]	[illegible]	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4	[illegible]
α[illegible]	ρ_1, ρ	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	.25	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	ρ_2, ρ	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.25	[illegible]	[illegible]	[illegible]
	ρ_{MS}, ρ	[illegible]			[illegible]			[illegible]		.25	[illegible]	[illegible]		[illegible]	1	[illegible]	[illegible]	[illegible]	4
α[illegible]	ρ_1, ρ	[illegible]	[illegible]			[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]	4	[illegible]		[illegible]	[illegible]	1	1	
	ρ_2, ρ					[illegible]	[illegible]	[illegible]	[illegible]		[illegible]	4	[illegible]		[illegible]	.25	1	[illegible]	[illegible]
	ρ_{MS}, ρ	[illegible]			[illegible]	4		[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	4	[illegible]	4		

Note 1. [illegible] δ [illegible] δ [illegible] δ [illegible]

Note 2. #n.bi: number of items with [illegible] entry, bias [illegible] and second entry, bias [illegible] blank or [illegible] number of items with [illegible] equals number of items with bias [illegible] min: smallest bias; max: largest bias [illegible]

Table 2

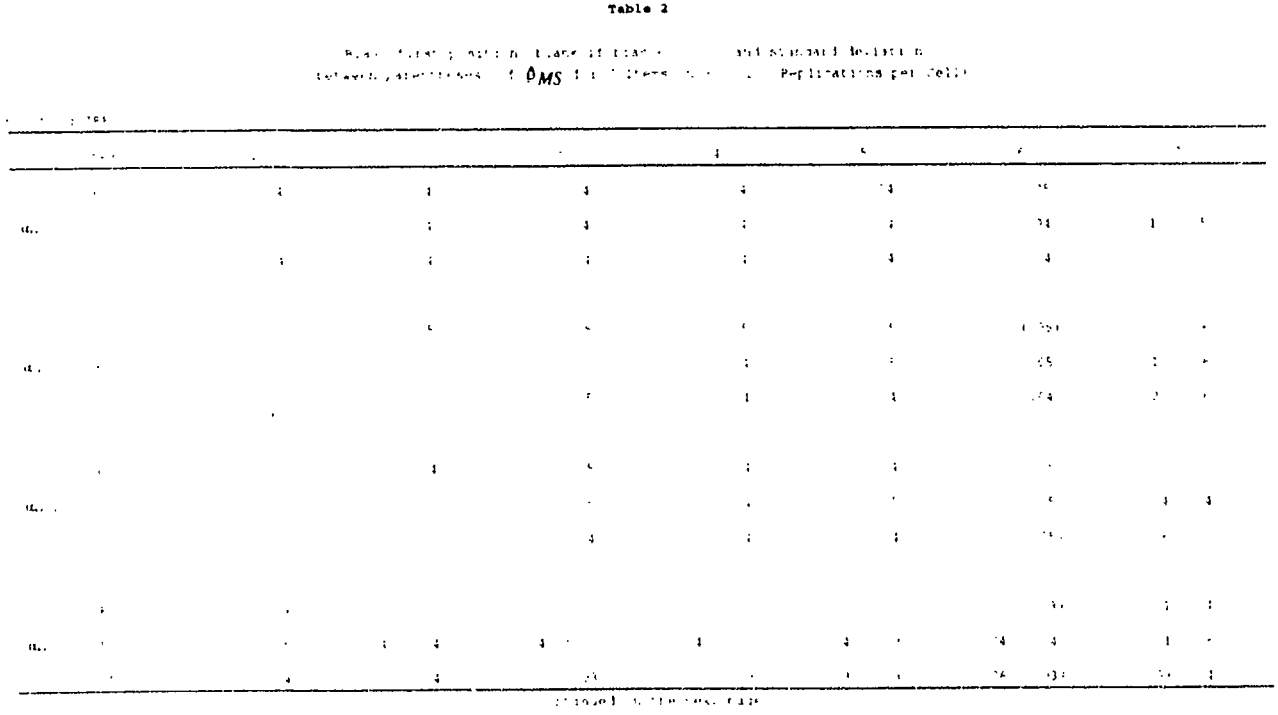


Table 3, continued

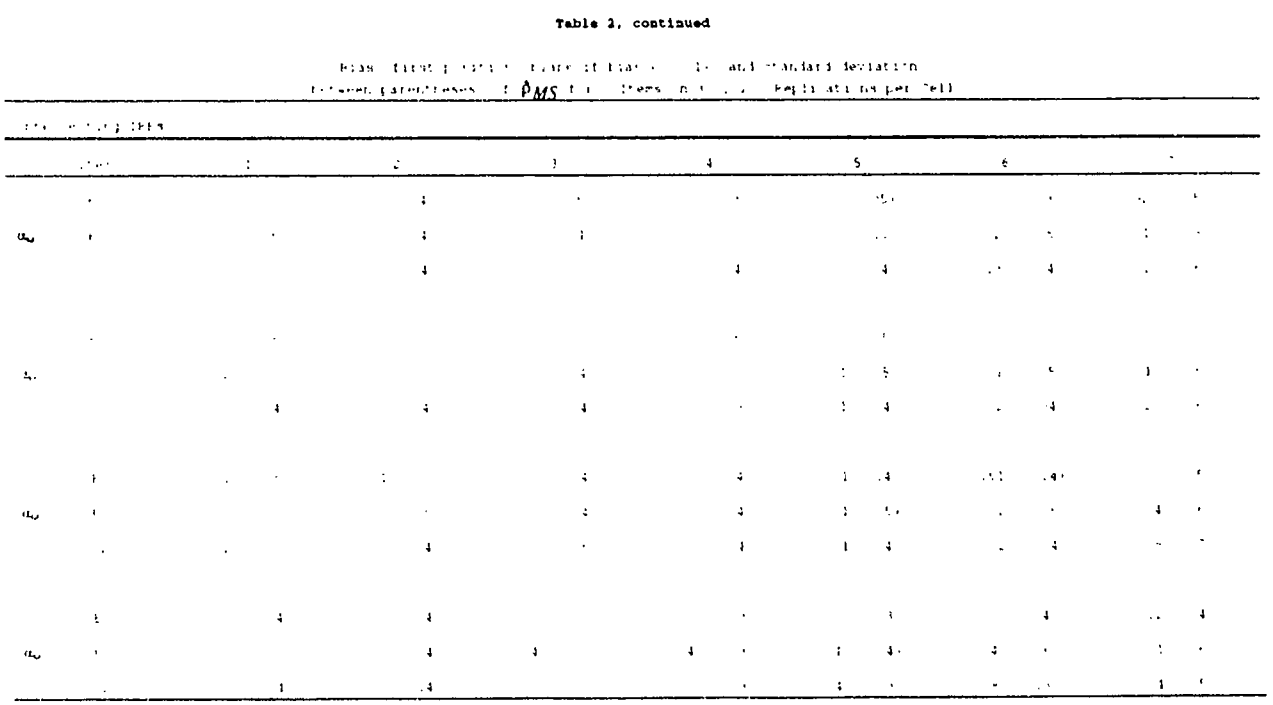
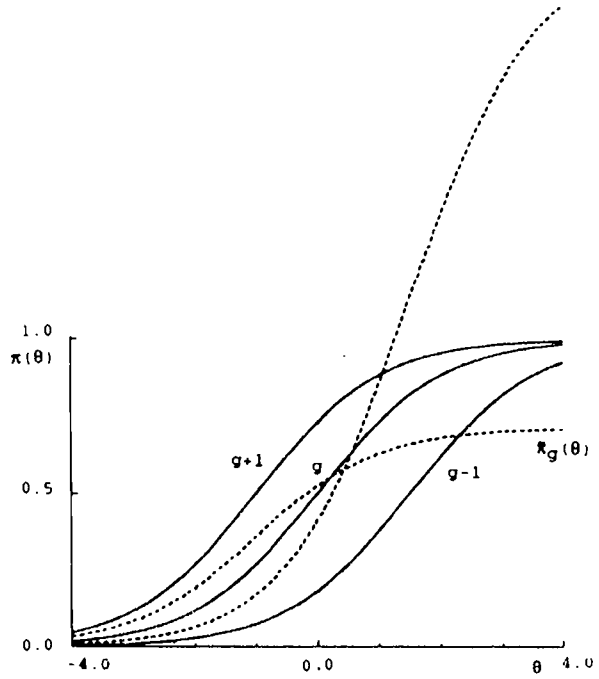
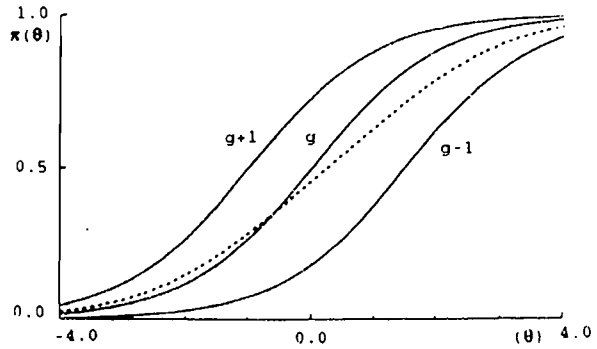


Figure Caption

Three IRFs with π_{g+1}=.697, π_g=.500, π_{g-1}=.222, $\pi_{g-1,g}$=.162, and $\pi_{g,g+1}$=.420, illustrating the approximation of $\pi_g(\Theta)$ by means of method 1 (left; dashed curves) and method 2 (right; dashed curves). Proportions based on δ_{g+1}=-1, δ_g=0, δ_{g-1}=1.5, α=1 for all three items and Θ standard normally distributed.





Titles of recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs from Bibliotheek, Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands