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

Rob R. Meijer, Universiteit Twente
Klaas Sijtsma, Rijksuniversiteit Utrecht
Ivo W. Molenaar, Rijksuniversiteit Groningen

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. 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 to the other two, because it has a smaller bias and a smaller sampling variance. This method also demonstrated some robustness under violation of the condition of nonintersecting item response functions.
Index terms: item reliability, item response theory, Mokken model, nonparametric item response models, test construction.

In practice, total scores on a test are more important than scores on individual items. In test construction, however, item quality must be assessed to select appropriate items that together will constitute a useful test. For example, in classical test theory (Lord & Novick, 1968) item statistics, such as the proportion correct 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, and pseudoguessing level. Moreover, the item information function (Lord, 1980, p. 72) can be used to assess measurement precision of a single item. The nonparametric Mokken approach to IRT (Mokken, 1971, in press; Mokken & Lewis, 1982) uses proportion correct and an item scalability coefficient.

Because the Mokken approach provides the theoretical framework for this study, its relevant assumptions and definitions are discussed. It is argued that in the Mokken IRT approach the reliability of an item can serve as a nonparametric counterpart of the item discrimination in logistic IRT and the corrected item-total correlation from classical test theory [refer to Lord (1980, p. 33) for a comparison of these latter two item statistics].

The purpose of this paper was 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 were investigated.

Basic Assumptions of the Nonparametric Mokken Approach

Nonparametric IRT models are important for ordering persons and items. Cliff & Donoghue (1992) provided arguments that favor ordinal rather than interval measurement in psychological and educational testing. Mokken (1971, pp. 115–169, in press; Mokken & Lewis, 1982) proposed two nonparametric IRT models for the analysis of binary item scores. The first was the monotone homogeneity model (MHM), which is defined by the assumptions of unidimensionality, local stochastic independence, and nondecreasingness of the item response functions

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 4, December 1995, pp. 323–335

0146-6216/95/040323-13$1.90

(IRFs). An important property of the MHM is that the latent trait score (θ) 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 & van den Wollenberg (1993), and Junker (1993); other ordinal models were investigated by Schulman & Haden (1975) and Cliff (1979).

The second model is the double monotonicity model (DMM). The DMM assumes unidimensionality, local stochastic independence, and nondecreasingness of the IRFs, as well as a fourth assumption that the IRFs do not intersect. Thus, the DMM 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 & Meijer (1992), and Sijtsma & Junker (in press).

The Rasch (1960) model is based on the three assumptions of the MHM, plus the fourth assumption of minimal sufficiency of the number-correct scores of persons and items for the estimation of θ and the 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) discussed conditions from which it can be derived that, in general, DMM IRFs cannot be transformed into Rasch IRFs. For example, the DMM allows IRFs with asymptotes that are not equal to 0 or 1, whereas the Rasch model excludes such IRFs. Disregarding the trivial case of constant IRFs, theoretically the DMM includes the Rasch model as a special case. In practice, however, differences become apparent for small numbers of items (e.g., at most 15 items). For larger numbers of items, the DMM still allows relatively easy items to have pseudoguessing 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 θ examinees, even if there is no guessing, and difficult items need not be trivial for high θ examinees.

Other differences between the DMM and the Rasch model are that DMM IRFs need not be symmetrical with respect to the inflection point and that the slopes of the IRFs may differ. The restriction that the IRFs are nondecreasing and nonintersecting implies that these variations can only be effective in short tests with item locations that are far apart. Meijer, Sijtsma, & Smid (1990) provided a theoretical and a practical comparison of the DMM and the Rasch model.

Because of the nonparametric definition of the IRFs, the MHM and the DMM do not assume specific distributions for latent model parameters; that is, characteristics of the models hold irrespective of such distributions. As a result of a nonparametric definition, latent item parameters from parametric models, such as item difficulty and discrimination, cannot be numerically estimated. In Mokken's (1971; Mokken & Lewis, 1982) nonparametric approach, item difficulty is replaced by the proportion of correct responses to 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, & Sijtsma (1986) noted that this coefficient is related to the slope of an IRF.

Item Reliability

Item Reliability and Repeatability

Donoghue & 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 discrimination than item scalability (Mokken, 1971, p. 151; Mokken & Lewis, 1982) 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, 1994; Tatsuoka & Tatsuoka, 1983; van der Flier, 1982). Here, item reliability is proposed as an appropriate replacement for item discrimination [also refer to Meredith (1965) for a similar proposal] in a nonparametric IRT context.

The reliability of an item expresses the degree to which observed item scores can be repeated independently under similar conditions. Item discrimination (denoted by α) as defined in logistic IRT (Lord, 1980, p. 13) has a similar interpretation. Let θ be the latent person parameter with probability

density $f(\theta)$. Furthermore, let item g ($g = 1, ..., k$) 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. As $\alpha_g \to \infty$, 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 examinee 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 probability correct is a constant irrespective of α_g. Thus, holding everything else constant, an increase in α_g corresponds to a higher degree of repeatability of item scores.

Definition and Estimation

Because the theoretical basis for the definition and the estimation of item reliability was given by Mokken (1971, pp. 142–147) and Sijtsma & Molenaar (1987), only results are provided here. Let π_g be the population proportion of persons giving a correct response on 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) defined the reliability of the dichotomous item score X_g as

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

$\rho(X_g) = 0$ if $\pi_{gg} = \pi_g^2$ (statistical independence between replications of item g); $\rho(X_g) = 1$ if $\pi_{gg} = \pi_g$.

π_g can be estimated unbiasedly (Mokken, 1971, p. 126); however, because locally independent replications of items are not possible, 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 & 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.

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

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

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 $\pi_{g-1}(\theta)$, $\pi_{g+1}(\theta)$, or both as a predictor of a real replication of item g. Note that π_{gg} equals

$$\pi_{gg} = \int_\theta \pi_g(\theta)\,\pi_g(\theta)\,dF(\theta), \quad (3)$$

where $F(\theta)$ is the cumulative distribution of θ.

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—$\pi_{g-1}(\theta)$, $\pi_{g+1}(\theta)$, or both:

$$\tilde{\pi}_g(\theta) = a + b\pi_{g-1}(\theta) + c\pi_{g+1}(\theta). \quad (4)$$

Each method is defined by the choice of a, b, and c. Substitution of $\tilde{\pi}_g(\theta)$ in Equation 3 and integration yield

$$\tilde{\pi}_{gg} = a\pi_g + b\pi_{g-1,g} + c\pi_{g,g+1}. \quad (5)$$

In Equation 5, $\pi_{g-1,g}$ 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 $\pi_{g,g+1}$:

$$\tilde{\pi}_{gg} = \frac{\pi_{g-1,g}\pi_g}{\pi_{g-1}} \text{ if } \pi_{g-1}(\theta) \text{ is used}, \quad (6)$$

and

$$\tilde{\pi}_{gg} = \frac{\pi_{g,g+1}\pi_g}{\pi_{g+1}} \text{ if } \pi_{g+1}(\theta) \text{ is used}. \quad (7)$$

Equation 6 should be used if π_{g-1} is closer to π_g than π_{g+1}; Equation 7 should be used otherwise. Sijtsma & Molenaar (1987) proposed a decision rule that resolves the problem of equal distances $\pi_g - \pi_{g-1} = \pi_{g+1} - \pi_g$.

Sijtsma & Molenaar (1987) provided the coun-

terparts for Equation 6 and Equation 7 after reversal of the scale direction—1s are replaced by 0s and 0s by 1s:

$$\tilde{\pi}_{gg} = \frac{\pi_{g-1,g}\left(1-\pi_g\right)}{1-\pi_{g-1}} + \frac{\pi_g\left(\pi_g-\pi_{g-1}\right)}{1-\pi_{g-1}}, \quad (8)$$

and

$$\tilde{\pi}_{gg} = \frac{\pi_{g,g+1}\left(1-\pi_g\right)}{1-\pi_{g+1}} - \frac{\pi_g\left(\pi_{g+1}-\pi_g\right)}{1-\pi_{g+1}}. \quad (9)$$

Because these four approximations are asymptotically biased with different signs (Molenaar & Sijtsma, 1984), Sijtsma & Molenaar's (1987) method used the unweighted mean of these four approximations, for which most of the bias cancels. This method is denoted by MS.

Mokken's Method 2 used both neighbors of item g to approximate π_{gg} by interpolation (Mokken, 1971, p. 147). The approximation formula (Mokken, 1971, p. 147) is

$$\tilde{\pi}_{gg} = \pi_{g-1,g} + \left(\pi_{g,g+1} - \pi_{g-1,g}\right)\frac{\pi_g - \pi_{g-1}}{\pi_{g+1} - \pi_{g-1}}. \quad (10)$$

For the two extreme items, extrapolation (Method 1) is used; refer to Sijtsma & Molenaar (1987) for further details. Substitution of an approximation to π_{gg} in $\rho(X_g)$ in Equation 1 yields ρ_1 if Equations 6 and 7 are used, ρ_2 if Equation 10 is used, and ρ_{MS} if the mean of Equations 6, 7, 8, and 9 is used. Note that Sijtsma & Molenaar (1987) only provided results pertaining to sample bias and variance of total score reliability estimation for each of the three reliability methods [Method 1 (Equations 6–7), MS (Equations 6–9), and Method 2 (Equation 10)].

All approximations to π_{gg} are functions of the bivariate proportions $\pi_{g-1,g}$ and $\pi_{g,g+1}$ and the distance between π_gs. If a bivariate proportion is smaller or a distance is larger than expected, compared with what it would have been if the items had been replications, this may bias $\tilde{\pi}_{gg}$ and, consequently, the reliability estimate of item g.

Illustration of Bias In the Methods

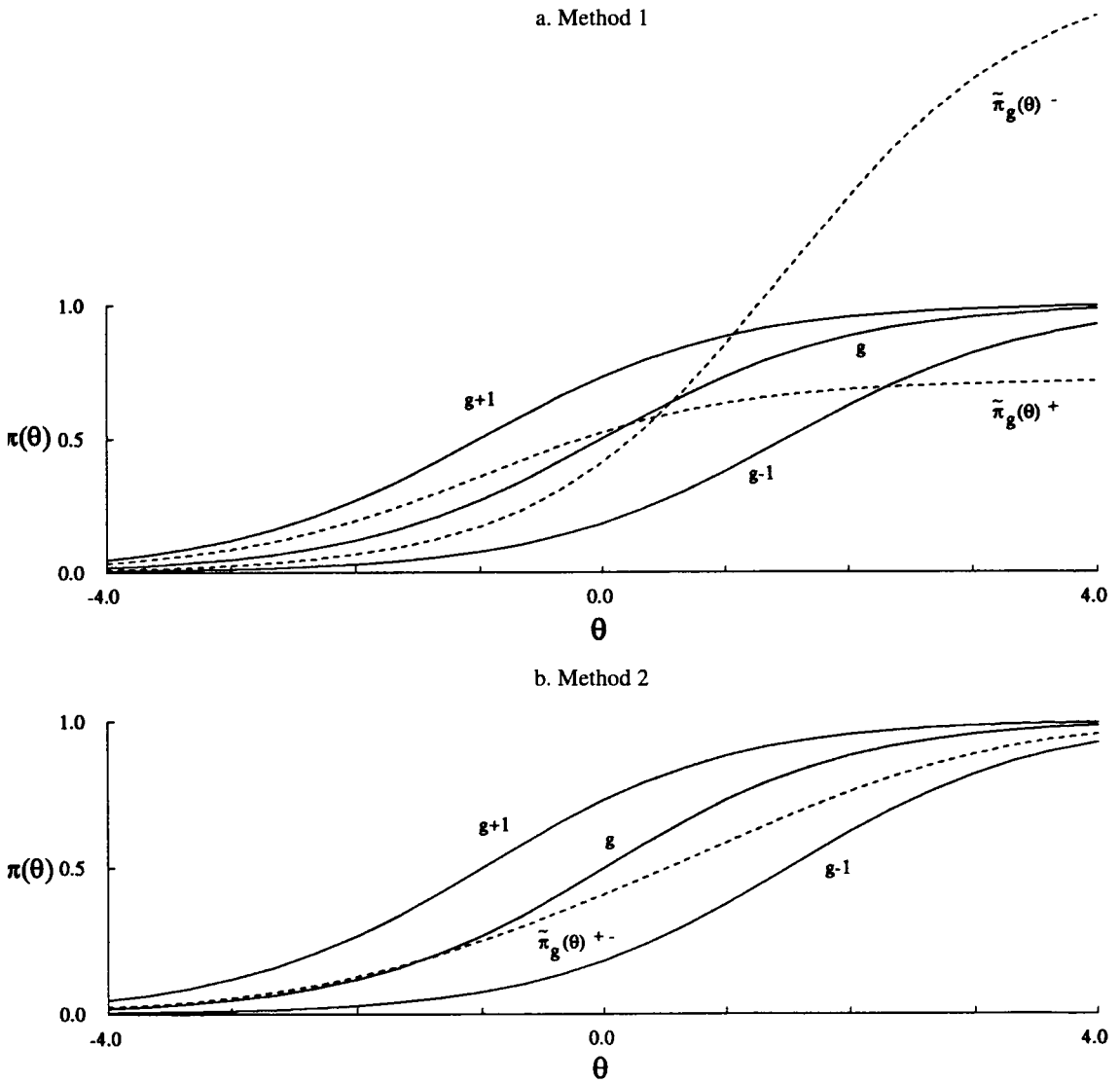
Figure 1 illustrates the effect of distance on the approximation of $\pi_g(\theta)$ using Method 1 (dashed curves in Figure 1a) and Method 2 (dashed curve in Figure 1b). In Figure 1, $\pi_{g+1} = .697$, $\pi_g = .500$, $\pi_{g-1} = .222$, $\pi_{g-1,g} = .162$, and $\pi_{g,g+1} = .420$. These proportions were based on $\delta_{g+1} = -1$, $\delta_g = 0$, $\delta_{g-1} = 1.5$, $\alpha = 1$, and θ was normally distributed. For Method 1, $\tilde{\pi}_g(\theta)^+$ was based on Equation 7, and $\tilde{\pi}_g(\theta)^-$ was based on Equation 6. For Method 2, $\tilde{\pi}_g(\theta)^{+-}$ was based on Equation 10.

To illustrate the estimation of π_{gg} using Method 1, the extrapolation formula given in Equation 7 is necessary. For Method 1, the dashed curve denoted $\tilde{\pi}_g(\theta)^+$ in Figure 1a is the approximation to $\pi_g(\theta)$ using $(\pi_g/\pi_{g+1})\pi_{g+1}(\theta)$ (note that substitution of this product in Equation 3 yields Equation 7). 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. Because 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 Equation 7. 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 and $\pi_g(\theta)$ is fixed, the multiplication factor π_g/π_{g+1} in Equation 7 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, $\pi_g(\theta)$ is more heavily underestimated if the distance is larger. The same line of reasoning leads to the conclusion that the approximation based on $\pi_{g-1}(\theta)$ [see Equation 6; also refer to the dashed line $\tilde{\pi}_g(\theta)^-$ in Figure 1a] 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 1b), 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 and keeping $\pi_g(\theta)$ and $\pi_{g+1}(\theta)$ fixed increases the inequality of the distances and leads to a situation in which it is difficult to predict how the bias of $\tilde{\pi}_{gg}$ will be affected.

These examples lead to the conclusion that distance affects the degree to which $\tilde{\pi}_{gg}$ is biased, and unequal distances of both neighbors to $\pi_g(\theta)$ affects

Figure 1
Three IRFs Illustrating the Approximation of $\pi_g(\theta)$ Using Methods 1 and 2



the bias differently than equal distances. Given the susceptibility of the item reliability methods to the quality of other items in the test, it was determined that it was important to investigate which of the three methods has the smallest bias.

An alternative approach would be to use 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, larger bias in estimating π_{gg} for $m > 2$ would be expected. 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 the test, items, and popula-

tion. [Donoghue & Cliff (1991) and Cliff & Donoghue (1992) used ordinal multiple regression for a related problem in ordinal true score theory.] Rather than pursuing a more complex strategy, asymptotic and sampling characteristics of reliability estimators based on the simpler Methods 1, 2, and MS were investigated. Only if none of these methods yields satisfactory results may a more complex strategy be rewarding.

In the simulation studies conducted here, the slopes of the IRFs were equal in several conditions, but this does not mean that the same reliability was estimated for each item. The estimation depends on: (1) the variance of the item score and, therefore, the proportion correct on the item; and (2) the difference between the proportions correct of the items used for the estimation. As a result, items with parallel IRFs can have different reliabilities although they have equal discriminations.

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

Asymptotic Bias in Item Reliability Estimation Methods

Method

The bias of the three item reliability methods was investigated with respect to $\rho(X_g)$ in Equation 1 using population fractions obtained from numerical integration across the θ distribution. This allowed the performance of the three methods to be investigated in the ideal case of very large samples.

Sets of seven items were used. Seven items was large enough because (1) the focus was on the individual item; (2) the distance between items could be manipulated equally well in small or large item sets; (3) the differences between extremely located items and items in between could be studied independently of test length; and (4) usually distances between adjacent items in longer tests are smaller so results for shorter tests were expected to be conservative. Logistic IRFs were used. Although the theoretical framework was nonparametric IRT, parametrically defined IRFs and parameter distributions were necessary to simulate 1s and 0s. However, this did not fully exploit the possibilities of the DMM. Such limitations are typical of research using simulated data in a nonparametric framework.

Given seven two-parameter logistic model (2PLM) IRFs and a standard normal distribution of θ, numerical integration (IMSL, 1987, routine QDAGS/DQDAGS) 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 ρ_1, ρ_2, and ρ_{MS}, respectively. The difference between each of these parameters and $\rho(X_g)$ is the bias of a specific method with respect to the reliability in Equation 1 for item g.

A completely crossed $4 \times 2 \times 3$ design was used. Four levels of average discrimination α_M were used: $\alpha_M = .5, 1, 2$, and 5. In combination with a standard normal θ distribution, these values ranged from very weak to very strong discrimination (Meijer et al., 1994).

Two levels of the spread of the αs within one test were used: no spread (all 7 αs equal) and positive spread (αs unequal). No spread corresponded to nonintersection of the 2PLM IRFs. For example, for $\alpha_M = 1$, $\alpha_g = 1$ ($g = 1, ..., 7$). Positive spread corresponded to intersection of the IRFs, and thus provided a violation of a condition underlying estimation of item reliability. For example, for $\alpha_M = 1$, $\alpha = (1.3, 1, 1, .7, 1, 1.3, .7)$. This more realistic condition allowed the robustness of the estimation methods to be investigated.

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. For two levels, item locations (δs) were equidistant with median of 0 and distance [$d(\delta)$] of either .1 or .5. These levels were denoted ES [equidistant, small distance (.1); $\delta = (-.3, -.2, -.1, 0, .1, .2, .3$ for the seven items)] and EL [equidistant, large distance (.5); $\delta = (-1.5, -1, -.5, 0, .5, 1, 1.5)$]. For the third level,

denoted UD (unequal distance), $d(\delta)$ varied more realistically within one item set. In particular, δ = (−.4, −.3, −.2, 0, .2, .8, 1.6) for all design cells at this level.

Results

Table 1 summarizes the asymptotic bias results for the complete design. For nonintersecting IRFs, the results for Method MS indicate that the reliability was almost unbiased for most items. For the ES, EL, and UD conditions taken together, 70 of the 84 reliabilities (12 cells with 7 items per cell) had an absolute bias smaller than .01, and 75 had an absolute bias smaller than .03. The largest bias was −.06 for $\alpha_M = 5$ and UD. The results were almost always better for Method MS than for Methods 1 and 2. Methods 1 and 2 often yielded unacceptably large absolute biases; for example, bias larger than .10. Method 1 often had a large bias for most of the 7 items in the test (not shown here). Method 2 primarily yielded large biases for the two extreme items (for which, in fact, Method 1 is used) and sometimes also for the items in between (also not shown here).

For intersecting IRFs, the asymptotic bias was larger for all three methods. For Method MS, 27 of the 84 reliabilities had an absolute bias smaller than .01, and 53 had an absolute bias smaller than .03. The largest bias for this method was −.07 for α_M = 5 and UD. As for the nonintersecting IRFs, the results for intersecting IRFs were almost always better for Method MS than for Methods 1 and 2. With a few exceptions, the bias of Method MS for single item reliabilities was acceptable (data for single items are not shown in Table 1).

For Method MS, a three-factor analysis of variance (ANOVA) was performed with bias as the dependent variable. Table 2 shows that with respect to the main effects, only α_M had a significant influence on the bias of items. Furthermore, there was only one significant two-way interaction between α_M and $d(\delta)$. The three-way interaction was not significant. Because no effect was found for the spread of the αs within a test, it was concluded that bias is quite robust given intersecting IRFs. Furthermore, the grand mean of the bias was .001 (not shown here). Main effects and interaction effects were almost all very close to 0 (between −.01 and .01), with one exception for $\alpha_M = 5$ and EL (first-order interaction was −.03; not shown here). For Method MS it can be concluded that: (1) bias was smaller for Method MS than for Methods 1 and 2; (2) bias was often negligible or practically acceptable; and (3) bias stayed within reasonable limits even if IRFs intersected.

Finite Sample Estimation of Item Reliability

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 Methods 1 and 2 (Table 1), they were included in the monte carlo investigation because (1) sampling variance as well as bias is important; (2) it may be that a method with larger asymptotic bias has smaller finite sample bias given, for example, 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 smaller.

Method

Data matrices containing binary item scores for N (persons) × 7 items were generated (for the simulation procedure see Sijtsma & Molenaar, 1987) using 2PLM 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; item δ and α values were the same as in the previous study. N = 100 was considered to be typical of ad hoc test construction that is part of a larger research project, N = 300 is typical of test construction research as performed in a noncommercial environment (e.g., universities where the means to collect data from larger samples are limited), and N = 900 (or more) is typical of large-scale test construction on a commercial basis.

Thus, a completely crossed 4 × 2 × 3 × 3 design was used. There were 200 replications in each cell. For each replication, the estimated π_g and π_{gh} were used (in the order found from that replicated data matrix) for estimation of ρ by Methods 1, 2, and MS.

Table 1
Number of Items With |Bias| < .01 and < .03, Largest Negative Bias (Min), and Largest Positive Bias (Max) for Parameters ρ_1, ρ_2, and ρ_{MS} Relative to $\rho(X_g)$, for ES, EL, and UD Conditions, and for Nonintersecting and Intersecting IRFs (Blank if |Bias| < .01)

Type of IRF, α_M, and Type of Bias	ES				EL				UD			
	Bias				Bias				Bias			
	.01	.03	Min	Max	.01	.03	Min	Max	.01	.03	Min	Max
Nonintersecting IRFs												
$\alpha_M = .5$												
$\rho_1 - \rho$	7	7			7	7			7	7		
$\rho_2 - \rho$	7	7			7	7			7	7		
$\rho_{MS} - \rho$	7	7			7	7			7	7		
$\alpha_M = 1$												
$\rho_1 - \rho$	7	7			1	3	−.04	.04	4	5	−.03	
$\rho_2 - \rho$	7	7			4	6	−.02	.04	5	6	−.03	
$\rho_{MS} - \rho$	7	7			5	7		.01	6	7		.02
$\alpha_M = 2$												
$\rho_1 - \rho$	1	7	−.03	.03	0	0	−.08	.14	1	5	−.12	.15
$\rho_2 - \rho$	5	7	−.02	.02	0	0	−.08	.14	3	5	−.12	.03
$\rho_{MS} - \rho$	7	7			5	6		.03	6	6		.05
$\alpha_M = 5$												
$\rho_1 - \rho$	0	0	−.06	.05	0	0	−.31	.25	0	1	−.40	.22
$\rho_2 - \rho$	5	5	−.05	.05	0	0	−.14	.25	1	3	−.40	.05
$\rho_{MS} - \rho$	7	7			2	2	−.04		4	5	−.06	
Intersecting IRFs												
$\alpha_M = .5$												
$\rho_1 - \rho$	1	1	−.04	.09	1	2	−.09	.05	1	3	−.03	.09
$\rho_2 - \rho$	2	3	−.12	.03	2	3	−.12	.05	2	3	−.12	.03
$\rho_{MS} - \rho$	2	5	−.06	.04	2	4	−.03	.05	2	4	−.03	.05
$\alpha_M = 1$												
$\rho_1 - \rho$	1	2	−.09	.06	0	2	−.09	.05	1	3	−.08	.04
$\rho_2 - \rho$	2	3	−.10	.06	1	3	−.09	.05	1	3	−.07	.04
$\rho_{MS} - \rho$	2	4	−.04	.05	2	4	−.03	.05	2	5	−.04	.05
$\alpha_M = 2$												
$\rho_1 - \rho$	1	3	−.05	.03	1	4	−.13	.18	1	2	−.07	.03
$\rho_2 - \rho$	2	3	−.05	.03	1	4	−.06	.15	1	3	−.09	.03
$\rho_{MS} - \rho$	2	5	−.04		2	4	−.01	.06	2	5	−.03	.04
$\alpha_M = 5$												
$\rho_1 - \rho$	1	3	−.04	.06	0	0	−.31	.25	1	1	−.40	.05
$\rho_2 - \rho$	2	5	−.04	.05	0	0	−.24	.25	1	1	−.40	.05
$\rho_{MS} - \rho$	4	7	−.01	.01	1	2	−.04	.01	4	4	−.07	

Results

For $N = 300$ and each combination of α_M, spread of the αs, and $d(\delta)$, the mean of an estimate of item reliability was calculated across items and 200 replications. Table 3 shows the mean finite sample bias and SD of the three methods in one test for $N = 300$. For both intersecting and nonintersecting IRFs, Method MS almost always had a smaller finite sample bias than Methods 1 and 2. Because it was so small, for practical purposes the bias of Method MS could be ignored. The SD of Method MS was almost always smaller than that of Methods 1 and 2. The same trend was found for $N = 100$ and $N = 900$. Because of these results, only the results for Method MS are discussed in more detail.

Nonintersecting IRFs. For nonintersecting IRFs, Table 4 (upper-half) shows that Method MS was almost unbiased. For the widely spaced items (EL) with nonintersecting IRFs, except for $\alpha_M = 5$, the bias was

Table 2
Results of the Analysis of Variance for Asymptotic Item Bias of ρ_{MS}

Source of Variation	Sum of Squares	*df*	*F*	*p*
Main Effects				
α_M	.015	3	9.018	0.000
$d(\delta)$	.000	2	.293	.746
Spread of α	.002	1	3.415	.067
Two-Way Interactions				
$\alpha_M \times d(\delta)$	.008	6	2.252	.042
$\alpha_M \times$ Spread of α	.001	3	.307	.821
$d(\delta) \times$ Spread of α	.001	2	.717	.490
Three-Way Interaction				
$\alpha_M \times d(\delta) \times$ Spread of α	.001	6	.232	.965
Error	.080	144		
Total	.107	167		

somewhat larger for the extreme items. For $\alpha_M = 5$, the bias was larger for nonextreme items. For unequally spaced items (UD), bias was negligible except for $\alpha_M = 1$, Item 7; $\alpha_M = 2$, Item 7; and $\alpha_M = 5$, Items 5–7. For $N = 100$ (not shown here), bias was in general somewhat higher, especially for $\alpha_M = .5$ and $\alpha_M = 1$. For $N = 900$ (also not shown here), bias results were highly similar to the results obtained for $N = 300$.

For $N = 300$ and nonintersecting IRFs, the SDs for almost all items were approximately .05. Only the SDs for the extremely easy and difficult items from widely spaced sets of items (EL) sometimes were somewhat larger (e.g., for Item 1, $\alpha_M = 1$, the SD was .07). For $N = 100$ (not shown here), the SD of Method MS across samples was rather large (between .07 and .13 for the extreme items and between .04 and .09 for the items in between). For $N = 900$ (also not shown here), the SD for almost all items was approximately .025. In general, for $N = 100$ the SD was approximately $\sqrt{3}$ times as large as for $N = 300$, and for $N = 900$ it was approximately $\sqrt{3}$ times as small as for $N = 300$.

For $\alpha_M = 1$ and $\alpha_M = 2$, the distribution of the MS estimator was rather symmetrical around its mean for all sample sizes (the skewness was between −.4 and .4). For $\alpha_M = .5$, the distribution was positively skewed for some items. For $\alpha_M = 5$ and all sample sizes, the distribution was negatively skewed for some items and positively skewed for others. The peakedness of the distribution was comparable to the normal distribution for all discrimination levels (in general, the kurtosis was approximately 3).

Intersecting IRFs. For intersecting IRFs (Table 4), the bias of Method MS was generally larger than for nonintersecting IRFs. The pattern of bias across

Table 3
Mean Absolute Bias of ρ_1, ρ_2, and ρ_{MS} Across 200 Replications and Seven Items, and Average SD Across Seven Items, for $N = 300$ (Blank if |Bias| < .01) for ES, EL, and UD Conditions and for Nonintersecting and Intersecting IRFs

α_M and ρ	Nonintersecting IRFs						Intersecting IRFs					
	ES		EL		UD		ES		EL		UD	
	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD
$\alpha_M = .5$												
ρ_1		.06		.06		.05	.05	.06	.05	.06	.05	.06
ρ_2		.05		.05		.04	.03	.06	.04	.07	.05	.05
ρ_{MS}		.04		.04		.04		.04	.01	.04	.01	.04
$\alpha_M = 1$												
ρ_1		.06	.03	.07	.04	.07	.05	.06	.06	.06	.04	.07
ρ_2		.05	.03	.06	.03	.06	.05	.05	.03	.06	.03	.07
ρ_{MS}		.05	.01	.05	.01	.05		.05		.05	.01	.05
$\alpha_M = 2$												
ρ_1	.02	.06	.03	.07	.04	.07	.03	.06	.07	.07	.08	.07
ρ_2	.02	.06	.05	.07	.07	.07	.03	.06	.06	.06	.05	.06
ρ_{MS}		.05	.01	.05	.01	.05		.04	.01	.05	.01	.05
$\alpha_M = 5$												
ρ_1	.04	.05	.14	.05	.20	.06	.04	.04	.20	.05	.25	.06
ρ_2	.03	.05	.13	.03	.14	.05	.05	.05	.18	.04	.23	.06
ρ_{MS}		.02	.03	.04		.02		.04	.03	.04	.03	.04

Table 4
Mean Bias and Standard Deviation (SD) Across Replications of $\hat{\rho}_{MS}$ For Seven Items (N = 300, 200 Replications Per Cell; Blank if |Bias| < .01) and for Nonintersecting and Intersecting IRFs

Type of IRF,	Item													
α_M, Distance	1		2		3		4		5		6		7	
Between δs	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Nonintersecting IRFs														
α_M = .5														
ES		.04		.04		.04		.04		.04		.05		.05
EL		.05		.04		.04		.04		.04		.04	.01	.05
UD		.04		.04		.04		.04		.04		.04		.05
α_M = 1														
ES		.05		.05		.05		.05		.05		.05		.06
EL	.01	.07		.05		.05		.04		.05		.05	.01	.06
UD		.05		.05		.05		.04		.04		.04	.02	.06
α_M = 2														
ES		.05		.04		.05		.04		.04		.05		.05
EL	.03	.07		.05		.05		.04		.05		.05	.04	.04
UD		.05		.05		.04		.04		.04		.05	.06	.07
α_M = 5														
ES		.04		.03		.03		.03		.03		.03	−.01	.04
EL	−.01	.06	−.04	.04	−.04	.03	.04	.03	.04	.03	−.04	.04	−.01	.06
UD		.04		.04		.03		.03	−.03	.03	−.06	.03	−.03	.04
Intersecting IRFs														
α_M = .5														
ES	−.02	.05		.04		.05		.03		.05		.03	.02	.05
EL	−.02	.06		.04		.04		.03		.03	.02	.05	.01	.03
UD	−.02	.05		.04		.05		.04		.04	.03	.04	.01	.06
α_M = 1														
ES		.05		.05		.05		.05		.05		.05		.05
EL	.01	.06		.05		.04		.05	.01	.05	.01	.05	.01	.05
UD		.04		.04		.04		.03	.01	.04	.02	.04	.02	.06
α_M = 2														
ES	−.01	.05	−.01	.05		.04		.04	−.01	.04	.01	.04		.05
EL	.02	.07		.05		.04		.04	.01	.05	.02	.06	.04	.06
UD	.02	.05		.04		.05		.04	−.01	.04	.02	.04	.05	.07
α_M = 5														
ES		.04		.04		.03		.03		.03		.04	−.02	.04
EL	−.02	.05	−.03	.04	−.04	.03	.04	.03	.04	.04	−.04	.06	−.01	.06
UD		.04		.04		.03		.03	−.04	.03	−.08	.03	−.04	.05

items within a test was rather inconsistent. Relatively few items showed a bias smaller than −.01 or larger than .01; bias for these items ranged from −.08 to .05. However, for the majority of the items bias was much smaller. The SD results for intersecting IRFs were comparable to those for nonintersecting IRFs. The same results were generally observed for skewness and kurtosis. Results for intersecting and nonintersecting IRFs were comparable across the three sample sizes.

ANOVA results. For Method MS, two four-factor ANOVAs (N = 300) were performed: one with bias as the dependent variable and one with SD as the dependent variable. Table 5 shows that α_M and the spread of the αs had a significant effect on the bias of the items. There was only one significant two-way interaction, between α_M and $d(\delta)$. No three- or four-way interactions were found to be significant. The grand mean of the bias was −.001. The vast majority of the absolute values of main and interaction effects were less than .01. A few exceptions occurred for some first-, second-, and third-order

interactions ($\alpha_M = 5$; effects between −.02 and .02 in most cases; never smaller than −.03 or larger than .03). Thus, the finite sample bias results were largely in agreement with the asymptotic bias results.

Table 5
Results of the Analysis of Variance for Finite Sample Bias of ρ_{MS} ($N = 300$)

Source of Variation	Sum of Squares	*df*	*F*	*p*
Main Effects				
α_M	.031	3	17.003	0.000
$d(\delta)$	0.000	2	.184	.832
Spread of α	.004	1	5.793	.017
N	.002	2	2.031	.133
Two-Way Interactions				
$\alpha_M \times d(\delta)$	.011	6	2.994	.007
$\alpha_M \times$ Spread of α	.001	3	.577	.630
$\alpha_M \times N$	.003	6	.848	.533
$d(\delta) \times$ Spread of α	0.000	2	.311	.733
$d(\delta) \times N$	0.000	4	.040	.997
Spread of $\alpha \times N$	.001	2	.490	.613
Three-Way Interactions				
$\alpha_M \times d(\delta) \times$ Spread of α	.001	6	.138	.991
$\alpha_M \times d(\delta) \times N$	.002	12	.232	.997
$\alpha_M \times$ Spread of $\alpha \times N$	.001	6	.324	.924
$d(\delta) \times$ Spread of $\alpha \times N$	0.000	4	.154	.961
Four-Way Interaction				
$\alpha_M \times d(\delta) \times$ Spread of $\alpha \times N$	.001	12	.156	1.000
Error	.264	432		
Total	.323	503		

Table 6 shows the results of the ANOVA using SD as the dependent variable. Note that $d(\delta)$ and N had a significant influence on the SD of the items. There was also a significant two-way interaction between α_M and $d(\delta)$. No three- or four-way interactions were significant.

Discussion

The estimation and use of item reliability is not a common practice in nonparametric test construction research. Moreover, like item reliability, the H_g coefficient (Mokken & Lewis, 1982) is an increasing function of the slope of the IRF (Mokken et al., 1986). Therefore, it would be interesting to compare these two item parameters. This topic is briefly discussed in relation to the H coefficient for two

Table 6
Results of the Analysis of Variance for the SD of ρ_{MS}

Source of Variation	Sum of Squares	*df*	*F*	*p*
Main Effects				
α_M	.002	3	2.559	.061
$d(\delta)$	.001	2	3.762	.024
Spread of α	0.000	1	.270	.604
N	.205	2	1,177.206	0.000
Two-Way Interactions				
$\alpha_M \times d(\delta)$	.001	6	2.418	.026
$\alpha_M \times$ Spread of α	0.000	3	.358	.783
$\alpha_M \times N$	.001	6	2.712	.074
$d(\delta) \times$ Spread of α	0.000	2	.061	.941
$d(\delta) \times N$	0.000	4	.513	.726
Spread of $\alpha \times N$	0.000	2	.250	.779
Three-Way Interactions				
$\alpha_M \times d(\delta) \times$ Spread of α	0.000	6	.265	.953
$\alpha_M \times d(\delta) \times N$	0.000	12	.369	.974
$\alpha_M \times$ Spread of $\alpha \times N$	0.000	6	.823	.552
$d(\delta) \times$ Spread of $\alpha \times N$	0.000	4	.163	.957
Four-Way Interaction				
$\alpha_M \times d(\delta) \times$ Spread of $\alpha \times N$	0.000	12	.244	.996
Error	.038	432		
Total	.249	503		

items, H_{gh}. This coefficient is defined (Mokken & Lewis, 1982) as

$$H_{gh} = \frac{\pi_{gh} - \pi_g \pi_h}{\pi_g (1 - \pi_h)} \text{ for } \pi_g \leq \pi_h . \quad (11)$$

Note that the covariance is in the numerator and the maximum possible covariance given π_g and π_h is in the denominator. The scalability coefficient for item g with respect to the other $k - 1$ items in the test, H_g, is defined by taking the sum across all covariances between item g and the other items in the numerator and the sum across all corresponding maximum covariances in the denominator (Mokken & Lewis, 1982).

There is remarkable equivalence between the item reliability of item g and the H coefficient for two independent replications of item g. Let X_g and $X_{g'}$ be two independent replications. Then

$$\text{Cov}(X_g, X_{g'}) = \pi_{gg} - \pi_g^2 . \quad (12)$$

The scalability coefficient for two independent rep-

lications thus equals

$$H_{gg} = \frac{\pi_{gg} - \pi_g^2}{\pi_g(1 - \pi_g)}. \quad (13)$$

This is exactly the reliability of item g: $H_{gg} = \rho(X_g)$. An interpretation of the item reliability, thus, is the scalability of an item with respect to an independent replication of that item.

This shows that item reliability is related to but not identical to the concept of scalability (Mokken & Lewis, 1982). The item reliability expresses how well item performance can be repeated under similar circumstances. The item scalability expresses the degree to which an item is scalable in the sense of the Guttman model together with the other $k - 1$ items in the test. To estimate $\rho(X_g)$, information is used from one or two neighbor items in the difficulty ordering that replace a real replication of item g. How to use $\rho(X_g)$ and H_g in a complementary way in a Mokken analysis is a topic for future investigation.

Method MS investigated here for estimating item reliability is also useful for obtaining information about the discrimination of the items. The method circumvents the identification problems that may arise in estimating the discrimination parameter in the two- and three-parameter logistic models. The one-parameter logistic model proposed by Verhelst & Glas (1995) is an attempt to estimate a hybrid model between the one- and two-parameter logistic models by imputing integer values for the αs, thereby avoiding the estimation problem. The fit of the model to the data and the usefulness of the estimates of the person parameters and the item difficulties are improved iteratively by adaptation of the imputations of the αs. Although Method MS does not provide the αs themselves, given the nonparametric context used here it may be used to obtain similar information.

Theoretically, item reliability provides information about the item that is independent of the other items in the test. In practice, one or two neighbor items are used to estimate this parameter. Item reliability does not provide information about the fit of an item in the Mokken model of double monotonicity but should be used as a practical index of the quality of items after model-data fit has been established. This is a common strategy in test construction using an IRT model: first a set of items is isolated that are in agreement with the theoretical requirements, and next some items may be removed that are not suited for practical use because, for example, their reliability is too low. Note in this context that IRT models, such as the one-, two-, and three-parameter logistic models and the model of double monotonicity, theoretically allow items with an almost 0 reliability (refer to Wood, 1978, for an example involving the fit of the Rasch model to random noise data). Such items, of course, do not have practical measurement value.

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 [Introduction to psychological test theory]. Bern: Huber.

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 NJ: 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. (1987). *User's manual stat/library IMSL.* Houston TX: Author.

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. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111–120.

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: de Gruyter.

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

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. (in press). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology.*

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.

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

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 215–237). New York: Springer.

Wood, R. (1978). Fitting the Rasch model—A heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

Author's Address

Send requests for reprints or further information to Rob R. Meijer, Universiteit Twente, Faculteit der Toegepaste Onderwijskunde, Vakgroep Onderwijskundige Meetmethoden en Data-analyse, P.O. Box 217, 7500 AE Enschede, The Netherlands. Internet: meijer@edte.utwente.nl.