PSYCHOMETRIKA—VOL. 52, NO. 1, 79–97
MARCH 1987

RELIABILITY OF TEST SCORES IN NONPARAMETRIC ITEM RESPONSE THEORY

KLAAS SIJTSMA

FREE UNIVERSITY OF AMSTERDAM, THE NETHERLANDS

IVO W. MOLENAAR

UNIVERSITY OF GRONINGEN, THE NETHERLANDS

Three methods for estimating reliability are studied within the context of nonparametric item response theory. Two were proposed originally by Mokken (1971) and a third is developed in this paper. Using a Monte Carlo strategy, these three estimation methods are compared with four "classical" lower bounds to reliability. Finally, recommendations are given concerning the use of these estimation methods.

Key words: Mokken scaling, nonparametric item response theory, reliability estimates, distributional properties.

The Doubly Monotone Model

Mokken (1971; also, see Henning, 1976; Mokken and Lewis, 1982; Stokman and Van Schuur, 1980) has presented an item response model that may be called nonparametric for two reasons: no mathematical form of the item characteristic function is specified and no assumptions are made concerning the distribution of the latent subject parameter. Consequently, when the model is fit empirically, items and subjects can only be partially ordered, in contrast to the well-known Rasch and Birnbaum models based on mathematically specified item characteristic functions.

Given a test consisting of dichotomously scored items, Mokken's model rests on four assumptions:

1. Item unidimensionality;
2. Local stochastic independence of item responses for a fixed subject;
3. Nondecreasing item characteristic functions, implying a monotone relation between the item success probabilities and the latent trait, ξ;
4. Identical ordering of success probabilities on a set of items for all values of ξ, that is, monotonicity in the item difficulties.

Because two kinds of monotonicity are distinguished, the model is called doubly monotone (Mokken, 1971). Item characteristic functions do not intersect, but they may touch and in the limiting case, may coincide. The reader is referred to Mokken and Molenaar (1982a, 1982b) for methods of empirically checking and testing the doubly monotone model.

Requests for reprints should be sent to Klaas Sijtsma, Vakgroep Arbeids- en Organisatiepsychologie, Free University, De Boelelaan 1081, 1081 HV Amsterdam, THE NETHERLANDS.

The authors are grateful for constructive comments from the reviewers and from Charles Lewis.

0033-3123/87/0300-7087$00.75/0

It is of interest to note that the Rasch model may be obtained by replacing the fourth assumption by the sufficiency of the unweighted raw scores of items and subjects for estimating the item difficulty parameter and subject parameter, respectively. Thus, the Rasch model may be considered a special case of the doubly monotone model where all item characteristic functions are "horizontally parallel," that is, they can be translated into each other along the horizontal ξ-axis. The two- and three-parameter models of Birnbaum as well as the normal ogive model of Lord are not special cases of the doubly monotone model, since they allow for intersecting item characteristic curves.

Based on the doubly monotone model, Mokken (1971) proposed two methods for estimating the reliability of test scores. In this study, they are critically discussed and improved, and a third method is developed. Across random samples, the bias and stability of these three methods are studied. With respect to these sampling properties, a comparison of the methods with four methods from classical test theory is also carried out. Before presenting the results from this study, we first briefly discuss measures used within the doubly monotone model to assess the quality of measurement. Second, the rationale is discussed for using reliability in nonparametric item response theory.

Quality Measures Within the Doubly Monotone Model

Mokken (1971) proposed two global measures for assessing the quality of measurement within the context of the doubly monotone model. The first is Loevinger's (1948) scalability coefficient, H, which evaluates the degree of Guttman homogeneity (Guttman, 1950; also, see Molenaar & Sijtsma, 1984). When empirical data conform reasonably well to the Guttman model, confidence may be placed in subjects' score patterns, thereby obtaining additional information besides simple raw scores. Recently, the role of H in Mokken's doubly monotone model has become the subject of debate (Jansen, 1982a, 1982b; 1983; Jansen, Roskam and Van den Wollenberg, 1982; Molenaar 1982b, 1982c; Sijtsma, 1984).

The second quality measure is the reliability coefficient from classical test theory. It will be shown in the subsequent sections of this paper how this coefficient can be estimated using the specific properties of the doubly monotone model, and to what extent such estimates, proposed by Mokken (1971, p. 142–147) and Molenaar and Sijtsma (1984), are superior to the classical estimates like KR-20.

The third quality aspect of a Mokken scale is not expressed by one coefficient but by a sequence of test procedures. They aim at establishing the correctness of the four assumptions listed in the previous section by examining some of their observable consequences. For such procedures the reader is referred to Mokken (1971), Mokken and Lewis (1982), and Molenaar (1982a, 1982b). It will be assumed in the sequel that the plausibility of the assumptions has been established prior to reliability estimation.

In the spirit of Lord (1980), reliability is defined by substituting the latent trait value, ξ, for the true score in the standard formulas. If random selection of a subject from a population is followed by random selection of an answer pattern according to the probabilities obtained from the item characteristic functions, the variance of the observed score X is decomposed as

$$\sigma^2(X) = \sigma^2_\xi[E(X \mid \xi)] + E_\xi[\sigma^2(X \mid \xi)]. \tag{1}$$

The first term is the true score variance and the second term the average error variation. Reliability is then defined by

$$\rho_{XX'} = \frac{\sigma^2_\xi[E(X \mid \xi)]}{\sigma^2(X)} = 1 - \frac{E_\xi[\sigma^2(X \mid \xi)]}{\sigma^2(X)}. \tag{2}$$

It will be shown how the right hand side of (2) can be estimated from the single and pairwise item popularities observed in a random sample of persons. Prior to this, a few remarks will be made about the usefulness of (2) in the context of the nonparametric item response model.

The Reliability Concept in the Doubly Monotone Model

Many publications on item response theory are rather critical about the classical reliability concept (e.g., Fischer, 1974; Lumsden, 1976; Samejima, 1977; Weiss & Davison, 1981). Its value depends strongly on the population considered. Moreover, in a parametric IRT model, the total score X is of less interest than the ability estimate, $\hat{\xi}$. The counterpart of (2) for the ability estimate, $\hat{\xi}$, is

$$\rho_{\hat{\xi}\hat{\xi}'} = 1 - \frac{E_{\xi}[\sigma^2(\hat{\xi} \mid \xi)]}{\sigma^2(\hat{\xi})}, \tag{3}$$

and is called the index of subject separation by Gustafsson (1977). Next, parametric IRT introduces the test information function, $I(\xi, \hat{\xi})$, which is asymptotically (for an infinite test length) equal to $1/\sigma^2(\hat{\xi} \mid \xi)$, for example, see Lord (1983). This function can be estimated from the data; for the Rasch model, Oosterloo (1984) gives the asymptotic properties of such estimates. For the use of the estimated test information function during the design stage, see Lord (1980). Note that Lord (1983) gives arguments for estimating reliability on the observed rather than on the latent scale.

Nonparametric IRT, with its weaker assumptions, does not lead to numerical estimates of subject or item parameters. They would be based on the assumption of a specific parametric class of item characteristic functions, and nonparametric IRT does not make this restriction. Lewis (1983) and Mokken and Lewis (1982) discuss a procedure to estimate a modified subject parameter. The present paper is restricted to the standard Mokken model, however. Its end product is an ordering of subjects based on total score, and an ordering of items according to observed item popularity. Apart from random error, the observed order of persons reflects the order with respect to ξ: the monotone relation between success probability and latent attribute implies

$$\xi_1 < \xi_2 \Rightarrow E(X \mid \xi_1) \leq E(X \mid \xi_2), \tag{4}$$

with strict inequality except in the rare case that all item characteristic functions are constant on the interval (ξ_1, ξ_2).

Consider a hypothetical independent replication of the test, leading to an observed score X'. A high value of $\rho_{XX'}$ defined by (2) implies that for most person pairs, $X_1 < X_2$ is equivalent to $X'_1 < X'_2$. It could be argued that Kendall's rank correlation, where each discordant pair indicates a nonreplicated ordering of two subjects, is more appropriate. This would deprive us, however, of the simple and highly meaningful decomposition (2) into within-ξ and between-ξ components. Computation of the overall probability of events like $(X_1 < X_2, X'_1 > X'_2)$ would require a specification of both the ability distribution and the item characteristic functions, and nonparametric IRT was designed to see how far one could get without such specifications. See Schulman and Haden (1975) for an alternative.

For any smooth and well spread frequency distribution of observed scores—which is desirable in many measurement contexts—the conclusions from rank and product moment correlations are almost equivalent. This leads us to prefer the latter—for which simple estimates based on our weak assumptions are available—to the former as an indicator of the overall replicability of the obtained ordering of subjects.

Wood (1978) has argued that a data matrix consisting of pure random noise passes

the customary goodness of fit tests for the Rasch model: This simply means that all persons have the same latent trait value, and all variation in observed score is due to error. The index of subject separation warns against such undesirable applications of the Rasch model. The reliability coefficient has precisely the same role in the nonparametric model, and is equally effective in assessing the extent to which the test reliably discriminates between individuals in the population of interest.

Fischer (1974, p. 133) and others have well summarized the drawbacks of the population dependent value of reliability as a quality index of the measurement instrument. There are some arguments, however, that support the thesis that such a population dependence is an advantage rather than a drawback. Suppose inspection of test data obtained from a sample drawn from a certain population leads to the conclusion that the Rasch model holds. A new investigation in a different population is envisaged. First of all, there is no compelling evidence that the responses from that population will also follow the Rasch model. Second, even if they do, the scores may well pile up on one end of the scale: not only is there no finite estimate for persons with a zero or perfect score, but also large clusters of persons with the same score inevitably lead to low discrimination. This implies that almost any investigator needs information on the quality of the test for discriminating individuals within the population to be measured: any population independent quantity or function is incomplete as an indicator of the success of the measurement effort.

There are instances where the standard error of measurement itself is of prime interest, for example, when a prediction for one individual is made from his/her test score. Given the arbitrary scale of test scores, however, even then a ratio of error variance to total observed variance might be more meaningful. This holds *a fortiori* in the more usual setting in which individuals are compared. Classical reliability, equivalent to one minus the average across persons of the just mentioned ratio, is a useful first order approximation of such a ratio for specific persons.

Summarizing, there are some good reasons for estimating the classical reliability when the nonparametric Mokken procedures are applied even though it does not tell the whole story.

Estimating Reliability According to Mokken

Mokken's approach to reliability estimation avoids the assumptions of equivalence that are usual in classical approaches. We give a short exposition of his methods and then propose some improvements and an alternative.

Letting $\pi_i(\xi)$ denote the success probability of a subject with latent trait value ξ on item i, the conditional variance of item i equals $\pi_i(\xi)[1 - \pi_i(\xi)]$, and the conditional variance of the raw test score, X, has the form

$$\begin{aligned}\sigma^2(X \mid \xi) &= \sum_i \pi_i(\xi)[1 - \pi_i(\xi)] \\ &= \sum_i [\pi_i(\xi) - \pi_i^2(\xi)]. \end{aligned} \tag{5}$$

The expectation of (5) across subjects yields

$$\begin{aligned}E_\xi[\sigma^2(X \mid \xi)] &= \int \sum_i [\pi_i(\xi) - \pi_i^2(\xi)]\, dG(\xi) \\ &= \sum_i (\pi_i - \pi_{ii}), \end{aligned} \tag{6}$$

where $G(\xi)$ is the cumulative distribution function of the latent trait, ξ. In (6), π_i denotes the proportion of subjects that respond positively to item i and π_{ii} denotes the proportion

of positive responses in two replications of item i. Substitution of (6) into the reliability formula (2) yields

$$\rho_{XX'} = 1 - \frac{\sum_i (\pi_i - \pi_{ii})}{\sigma^2(X)}. \tag{7}$$

In (7), π_i can be estimated by means of the unbiased and consistent estimator, $\hat{\pi}_i = n_i/n$, where n_i denotes the number of positive responses to item i and n the total number of responses to item i. The situation is similar for the proportion of subjects answering positively to both item i and j ($i \neq j$), where $\hat{\pi}_{ij}$ is an estimator of

$$\pi_{ij} = \int \pi_i(\xi)\pi_j(\xi)\, dG(\xi). \tag{8}$$

We will use the proportions π_i and π_{ij} ($i \neq j$) to obtain an approximation of π_{ii}. Analogously to π_{ij} ($i \neq j$), π_{ii} can be written as

$$\pi_{ii} = \int \pi_i(\xi)\pi_i(\xi)\, dG(\xi). \tag{9}$$

Next, one of the factors $\pi_i(\xi)$ in the integrand is approximated by a linear function of one (Mokken's Method 1) or two (Mokken's Method 2) probabilities $\pi_j(\xi)$:

$$\tilde{\pi}_i(\xi) = a + b\pi_j(\xi) + c\pi_h(\xi), \qquad j \neq i, \qquad h \neq i, \tag{10}$$

where $\tilde{\pi}_i(\xi)$ denotes the approximation of $\pi_i(\xi)$. Inserting (10) for one of the factors $\pi_i(\xi)$ in (9) and integrating, yields an approximation, $\tilde{\pi}_{ii}$, to π_{ii}:

$$\tilde{\pi}_{ii} = a\pi_i + b\pi_{ij} + c\pi_{ih}, \tag{11}$$

where $\tilde{\pi}_{ii}$ can be estimated from empirical data because all probabilities on the right hand side have natural estimates. Mokken uses two different linear combinations for approximating π_{ii} that constitute his methods 1 and 2 for reliability estimation.

Double monotonicity of item characteristic functions implies, when items are ordered from difficult to easy,

$$\pi_{i-1}(\xi) \leq \pi_i(\xi) \leq \pi_{i+1}(\xi). \tag{12}$$

In Mokken's Method 1

$$\tilde{\pi}_i(\xi) = \pi_{i-1}(\xi)\frac{\pi_i}{\pi_{i-1}}, \qquad \text{implying} \qquad \tilde{\pi}_{ii} = \pi_{i-1,i}\frac{\pi_i}{\pi_{i-1}}, \tag{13}$$

and

$$\tilde{\pi}_i(\xi) = \pi_{i+1}(\xi)\frac{\pi_i}{\pi_{i+1}}, \qquad \text{implying} \qquad \tilde{\pi}_{ii} = \pi_{i,i+1}\frac{\pi_i}{\pi_{i+1}}. \tag{14}$$

According to Mokken (1971, p. 147) one should choose (13) if π_i is closer to π_{i-1} than to π_{i+1}, and (14) otherwise. When a test consists of k items, then for $i = 1$ and $i = k$, only one possible choice exists.

In Mokken's Method 2, $\tilde{\pi}_i(\xi)$ is an interpolation of $\pi_{i-1}(\xi)$ and $\pi_{i+1}(\xi)$:

$$\tilde{\pi}_i(\xi) = \pi_{i-1}(\xi) + [\pi_{i+1}(\xi) - \pi_{i-1}(\xi)]\frac{\pi_i - \pi_{i-1}}{\pi_{i+1} - \pi_{i-1}}, \tag{15}$$

implying

$$\tilde{\pi}_{ii} = \pi_{i-1,i} + [\pi_{i,i+1} - \underline{\pi}_{i-1,i}] \frac{\pi_i - \pi_{i-1}}{\pi_{i+1} - \pi_{i-1}}. \tag{16}$$

Mokken recommends Method 1 for $i = 1$ and $i = k$.

Molenaar and Sijtsma (1984) have compared Method 1, Method 2 and the classical reliability estimate, coefficient alpha (Cronbach, 1951), in a mathematically defined population. Numerical integration yielded the population values; no item responses, either empirical or simulated, were used in this study. Two-parameter logistic functions for $\pi_i(\xi)$ and a standard normal distribution for ξ were used. It turned out that Mokken's methods mostly led to only slightly biased approximations of the true reliability. Furthermore, they were always less biased than coefficient alpha. So, it seems worth while to study the sampling behavior of Mokken's reliability estimates and compare them with some classical lower bounds to reliability. Before doing so, we will propose a new method for estimating reliability within the context of the doubly monotone model, as well as some theoretical improvements for estimating π_{ii}.

A New Method for Reliability Estimation

Molenaar and Sijtsma (1984) proposed a new method for estimating reliability that promises to lead to less biased estimates than Mokken's Methods 1 and 2. Before presenting this method, however, two additional approximations to π_{ii} must be discussed that apply when the scale direction is reversed, that is, when negative and positive item scores are interchanged. The proportion of "ones" now equals $1 - \pi$. Substitution into formulas (15) and (16) of Method 2 yields the same results, but substitution into formulas (13) and (14) of Method 1 yields

$$\tilde{\pi}_{ii} = \pi_{i-1,i} \frac{1 - \pi_i}{1 - \pi_{i-1}} + \pi_i \frac{\pi_i - \pi_{i-1}}{1 - \pi_{i-1}}, \tag{17}$$

and

$$\tilde{\pi}_{ii} = \pi_{i,i+1} \frac{1 - \pi_i}{1 - \pi_{i+1}} - \pi_i \frac{\pi_{i+1} - \pi_i}{1 - \pi_{i+1}}, \tag{18}$$

respectively. This result is somewhat surprising since reversal of the scale does not influence the population reliability; Mokken's Method 1 leads to different results, however, when applied to original and reversed scales.

In the analytical study mentioned earlier, Molenaar and Sijtsma (1984) obtained indications that π_{ii}, when estimated by means of (13) and (18), tends to be negatively biased, while estimates obtained by means of (14) and (17) tend to be positively biased. When averaged, the biases more or less cancel each other, implying that a less biased estimate of π_{ii} can be obtained by simply taking the mean of the four estimates mentioned. This suggestion comprises the third method of estimating π_{ii} and reliability. The population results in Molenaar and Sijtsma mostly favor this third method in terms of bias compared to Mokken's methods (without the refinements proposed in the next section, however).

Refinements Concerning the Reliability Estimates

Methods 1 and 2 by Mokken and the method proposed by Molenaar and Sijtsma are not always applicable to empirical data. For example, when three or more item difficulties (proportions correct) are equal in the sample, for example, $\hat{\pi}_{i-1} = \hat{\pi}_i = \hat{\pi}_{i+1}$, it

is impossible to approximate $\pi_{i-1,i-1}$, π_{ii} and $\pi_{i+1,i+1}$ by means of Method 2 (see (16)) since the items can not be ordered. Consequently, for a specific item we can not decide which items are its neighbors in the ordering of item difficulties. Solutions for this and similar problems are proposed in the sequel.

Before doing so, we will say something about the arrangement of this section. The five approximation methods to π_{ii} are arranged vertically in Table 1, where they are denoted by the corresponding formula numbers. In several instances, the estimation of π_{ii} by some specific formula may be problematic. Since the choice in Mokken's Method 1 may also pose a problem, this method is depicted separately in Table 1. The problems that may occur in practice when estimating π_{ii} are arranged horizontally in Table 1. In the following subsections we will subsequently explain these problems and propose solutions, that is, we will treat Table 1 by columns.

Upper and Lower Bounds to π_{ii}

Approximations to π_{ii} may sometimes become unreasonably low or high, and therefore, we will derive bounds for π_{ii}. As a lower bound, we propose the instance of global independence among replications of item i, in which case $\pi_{ii}(L) = \pi_i^2$. Values of π_{ii} below this bound are regarded as unreasonable, since they would imply a negative correlation beween replications. It should be noted that negative correlations between items or replications are not allowed in the doubly monotone model (Mokken, 1971, pp. 120, 130). As an upper bound to π_{ii}, we propose $\pi_{ii}(U) = \pi_i$, which is a rather evident choice. Whenever the $\tilde{\pi}_{ii}$ are estimated empirically, say by $\hat{\tilde{\pi}}_{ii}$, the condition $\hat{\pi}_i^2 \leq \hat{\tilde{\pi}}_{ii} \leq \hat{\pi}_i$ should be checked. When $\hat{\tilde{\pi}}_{ii}$ is outside the interval, it should be replaced by the appropriate bound. (In the sequel we will write $\hat{\pi}_{ii}$ instead of $\hat{\tilde{\pi}}_{ii}$ for notational convenience).

Methods

All the methods for approximating π_{ii} that are listed vertically in Table 1 may yield estimates that are smaller than the lower bound $\pi_{ii}(L)$. With the exception of (16), which is Mokken's Method 2, $\tilde{\pi}_{ii} < \pi_{ii}(L)$ if and only if the correlation between the items involved in the formula is negative. As an example in the case of (13), solving the inequality

$$\pi_{i-1,i}\frac{\pi_i}{\pi_{i-1}} - \pi_i^2 < 0,$$

yields the condition $\pi_{i-1,i} < \pi_{i-1}\pi_i$, which implies a negative correlation between items $i-1$ and i.

Negative correlations between items are not allowed in the doubly monotone model. In empirical applications no serious problem arises since one usually starts the item analysis by removing negatively correlated items from the initial item set.

In the case of approximation (16), which is Mokken's Method 2, it is possible to obtain values of $\tilde{\pi}_{ii}$ smaller than $\pi_{ii}(L)$. As a numerical example, consider $\hat{\pi}_{i-1,i} = \hat{\pi}_{i,i+1} = 0$, yielding $\hat{\pi}_{ii} = 0$ according to approximation (16). Another example, with $\hat{\pi}_{i-1,i} = 0$, $\hat{\pi}_{i,i+1} = .20$, $\hat{\pi}_{i-1} = .20$, $\hat{\pi}_i = .25$ and $\hat{\pi}_{i+1} = .70$, yields $\hat{\pi}_{ii} = .0200$ while $\pi_{ii}(L) = .0625$. Analytically, no easily interpretable conditions could be obtained under which $\tilde{\pi}_{ii} < \pi_{ii}(L)$.

It can be shown that none of the methods for the approximation of π_{ii} can yield values that are larger than the upper bound $\pi_{ii}(U)$.

The results of this subsection are summarized in the first two columns of Table 1.

Equality of Univariate Proportions Correct

Equality of univariate proportions of correct answers poses problems to the appli-

TABLE 1

A "plus" sign means that a problem (horizontal) is valid for a method for estimating π_{ii} (vertical), while a "minus" sign either denotes the contrary or implies that a problem is irrelevant to the method under consideration.

	lower bound to π_{ii}: $\hat{\pi}_{ii} < \pi_i^2$	upper bound to π_{ii}: $\hat{\pi}_{ii} > \pi_i$	m equal proportions correct: $m \geq 2$	m equal proportions correct: $m \geq 3$	equidistance of proportions correct: $\pi_{i+1} - \pi_i = \pi_i - \pi_{i-1}$
(13)	+	-	+	-	-
(14)	+	-	+	-	-
method 2: (16)	+	-	-	+	-
(17)	+	+	+	-	-
(18)	+	-	+	-	-
method 1: (13),(14)	+	-	-	+	+

cation of all methods that are vertically listed in Table 1. Before considering these, we prove two theorems with respect to equal proportions correct in the doubly monotone model.

In the doubly monotone model the following two theorems hold true in the population:

Theorem 1.

$$\pi_{i+1}(\xi) = \cdots = \pi_{i+m}(\xi) \quad \text{iff} \quad \pi_{i+1} = \cdots = \pi_{i+m}. \tag{19}$$

Proof (necessity). Consider two items i and j:

$$\pi_i = \int \pi_i(\xi)\, dG(\xi) = \int \pi_j(\xi)\, dG(\xi) = \pi_j.$$

Proof (sufficiency).

$$\pi_i = \pi_j \Rightarrow \int \pi_i(\xi)\, dG(\xi) = \int \pi_j(\xi)\, dG(\xi).$$

In the doubly monotone model, where $\pi_i(\xi) \leq \pi_j(\xi)$ for all values of ξ $(i < j)$, this can only be true when $\pi_i(\xi) = \pi_j(\xi)$ for all values of ξ. Generalization of the theorem to m items having equal proportions of positive answers follows easily. □

Theorem 2.

$$\pi_{i,i+1} = \cdots = \pi_{i,i+m} = \pi_{i+1,i+2} = \cdots = \pi_{i+m-1,i+m} \quad \text{iff} \quad \pi_i = \cdots = \pi_{i+m}.$$

Proof (necessity). Consider just two bivariate proportions correct and assume equality: $\pi_{i,i+1} = \pi_{i,i+2}$, or:

$$\int \pi_i(\xi)\pi_{i+1}(\xi)\, dG(\xi) = \int \pi_i(\xi)\pi_{i+2}(\xi)\, dG(\xi). \tag{20}$$

This equality can be rewritten as

$$\int \pi_i(\xi)[\pi_{i+1}(\xi) - \pi_{i+2}(\xi)]\, dG(\xi) = 0. \tag{21}$$

Excluding the trivial solution $\pi_i(\xi) = 0$ for all ξ, in the context of the doubly monotone model, where $\pi_{i+1}(\xi) \leq \pi_{i+2}(\xi)$ for all ξ, the only solution to (21) is $\pi_{i+1}(\xi) = \pi_{i+2}(\xi)$, implying $\pi_{i+1} = \pi_{i+2}$. This result can easily be generalized to the case of m items.

Proof (sufficiency). Consider two proportions correct and assume equality: $\pi_i = \pi_{i+1}$. That is,

$$\int \pi_i(\xi)\, dG(\xi) = \int \pi_{i+1}(\xi)\, dG(\xi),$$

which can be written as

$$\int [\pi_i(\xi) - \pi_{i+1}(\xi)]\, dG(\xi) = 0.$$

Since the double monotone condition holds, $\pi_i(\xi) \leq \pi_{i+1}(\xi)$ for all ξ, and thus equality holds for all ξ. This result can be generalized to the case of m items. Substitution of the general result in (8) yields equality of the corresponding bivariate proportions. □

Two important results follow from our derivations. Combining Theorem 1 and (9) for the arbitrary case of three items i, j and h, leads to the conclusion that $\pi_{ii} = \pi_{jj} = \pi_{hh}$.

Also, $\pi_{ij} = \pi_{ih} = \pi_{jh}$ and it can easily be inferred that $\pi_{ii} = \pi_{ij}$, and so forth.

Population methods. In the case of two or more items having equal proportions correct ($m \geq 2$; m denotes the number of items with equal proportions correct), application of (13), (14), (17) and (18) leads to different approximations of π_{ii} for such items. This result is contrary to our theory, and consequently, other approximation methods should be devised.

As an example, assume $m = 2$ and consider four items called A, B, C and D, with two possible orderings of the proportions correct:

I. $\pi_A = .3, \pi_B = .5, \pi_C = .5, \pi_D = .6$;
II. $\pi_A = .3, \pi_C = .5, \pi_B = .5, \pi_D = .6$.

Consecutively, we replace their names by increasing rank numbers that are assigned according to the items' order from left to right.

Using (17) for approximating π_{BB} yields

$$\text{I. } \tilde{\pi}_{22} = \pi_{12} \frac{1 - \pi_2}{1 - \pi_1} + \pi_2 \frac{\pi_2 - \pi_1}{1 - \pi_1},$$

where item B has rank 2, and

$$\text{II. } \tilde{\pi}_{33} = \pi_{23},$$

where item B has rank 3.

Using (18) for approximating π_{BB} yields

$$\text{I. } \tilde{\pi}_{22} = \pi_{23}, \quad \text{and}$$

$$\text{II. } \tilde{\pi}_{33} = \pi_{34} \frac{1 - \pi_3}{1 - \pi_4} - \pi_3 \frac{\pi_4 - \pi_3}{1 - \pi_4},$$

respectively. Similar results can be obtained from (13) and (14).

The case of $m = 2$ does not provide a problem for Mokken's Methods 1 and 2, respectively. This can be shown by making up similar examples as we have done for the other approximation methods.

Finally, we consider $m = 3$ and five items, where

$$\pi_A = .2, \quad \pi_B = .5, \quad \pi_C = .5, \quad \pi_D = .5, \quad \pi_E = .6.$$

Again, replacing names by rank numbers and considering only the ordering just given, we approximate π_{BB}, π_{CC} and π_{DD} by means of Mokken's Method 1:

$$\tilde{\pi}_{22} = \pi_{23},$$

$$\tilde{\pi}_{33} = \pi_{23} \quad \text{or} \quad \tilde{\pi}_{33} = \pi_{34}, \quad \text{and}$$

$$\tilde{\pi}_{44} = \pi_{34}.$$

In the case of Method 2 or (16):

$$\tilde{\pi}_{22} = \pi_{23},$$

$$\tilde{\pi}_{33} = \pi_{23} + (\pi_{34} - \pi_{23}) \frac{\pi_3 - \pi_2}{\pi_4 - \pi_2},$$

which is undefined, and

$$\tilde{\pi}_{44} = \pi_{34}.$$

These results are also contrary to our theory.

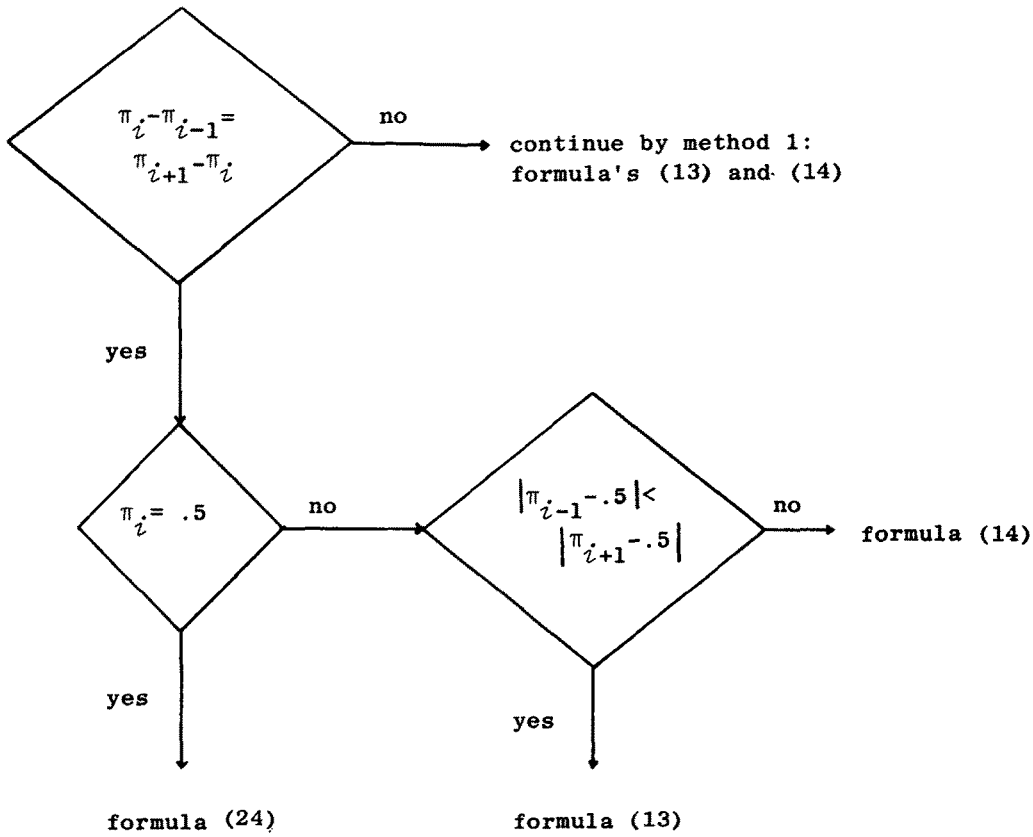

FIGURE 1
Decisions to be taken when $\pi_i - \pi_{i-1} = \pi_{i+1} - \pi_i$ in Mokken's Method 1.

An estimation method for π_{ii}. After insertion of sample proportions, this subsection has only considered population results. In the sample, we assume a set, C, consisting of m items having equal observed univariate proportions correct. From our theory, we may estimate π_{ii} by means of $\hat{\pi}_{ij}$, where i and j belong to C. But actually, π_{ii} can be estimated by any bivariate proportion $\hat{\pi}_{jh}$, where $j \neq h$ and j and h belong to C. Then, we can improve the estimation of π_{ii} by using all bivariate proportions of the m items belonging to C:

$$\hat{\pi}_{ii} = \frac{2}{m(m-1)} \sum_{\substack{j<h \\ j,h \in C}} \hat{\pi}_{jh}. \tag{22}$$

This estimate can be used for all items belonging to C.

Equal Distances Between Univariate Proportions Correct

The third problem occurs only in Method 1 in the sample when $\hat{\pi}_i - \hat{\pi}_{i-1} = \hat{\pi}_{i+1} - \hat{\pi}_i$, since no choice can be made whether to use (13) or (14). A solution might be based on the sampling variability of proportions, $\sigma^2(\hat{\pi}) = \pi(1-\pi)/n$, which is at its maximum when $\pi = .5$ and decreases towards the extremes. When we take, for instance, $\hat{\pi}_{i-1} = .5$, $\hat{\pi}_i = .6$ and $\hat{\pi}_{i+1} = .7$, we may argue that since the confidence interval for π_{i-1} is larger than the one for π_{i+1} the difference $\pi_i - \pi_{i-1}$ is likely to be smaller than

$\pi_{i+1} - \pi_i$ when one is willing to take the confidence bounds into consideration. The decision would then be to choose (13). For the example, $\hat{\pi}_{i-1} = .10$, $\hat{\pi}_i = .35$ and $\hat{\pi}_{i+1} = .60$ we would choose (14) following the same rationale. So, in equidistance cases one could always approximate the item characteristic function with item proportion correct that is closest to .5.

When $\hat{\pi}_i = .5$ no clearcut decision follows from our rule and we will resort to Method 2, which in that case for population results is adapted to

$$\tilde{\pi}_i(\xi) = \pi_{i-1}(\xi) + \tfrac{1}{2}[\pi_{i+1}(\xi) - \pi_{i-1}(\xi)], \tag{23}$$

where the multiplication constant equals one half so that equal use is made of both item characteristic functions. Finally, the estimate of π_{ii} now becomes

$$\hat{\pi}_{ii} = \tfrac{1}{2}(\hat{\pi}_{i-1,i} + \hat{\pi}_{i,i+1}). \tag{24}$$

The decisions that have to be taken when $\hat{\pi}_i - \hat{\pi}_{i-1} = \hat{\pi}_{i+1} - \hat{\pi}_i$ are depicted in Figure 1.

Sampling Behavior of the Reliability Estimates

Molenaar and Sijtsma (1984) have studied in the population the reliability approximations of Mokken (1971) and a third approximation using the average of (13), (14), (17) and (18) presented in this paper, denoted by "MS" in the tables. What matters especially is the behavior of these approximations in samples, which can be studied in two ways.

First, we may try to derive analytically test statistics with a known sampling distribution. Unfortunately, there is an a priori reason why this is very hard and perhaps impossible to accomplish. When two items have population difficulties that are close together, reversals of their order may be expected to occur in a nontrivial number of samples drawn from the same population. Consequently, different approximation methods to π_{ii} may be prescribed in different samples, since a fixed item may have different neighbors in different samples. In short, it is not to be expected on a priori grounds that theoretical assertions with respect to distributional properties of the reliability estimates will hold good in empirical samples.

Second, as we discuss below, we may actually study the behavior of the reliability estimates across a large number of random samples from the same population in a Monte Carlo study.

Statistics of Interest

Whether the reliability estimates discussed in this paper are better estimates than already existing estimates with respect to stability and bias across random samples of subjects is an open question. From the large number of existing reliability estimates (see e.g., Cronbach, 1951; Guttman, 1945; Jackson and Agunwamba, 1977; ten Berge & Zegers, 1978; ten Berge, Snijders & Zegers, 1981), an obvious choice is the lower bound to $\rho_{XX'}$, known as coefficient alpha (Cronbach, 1951).

Ten Berge and Zegers (1978) have presented an infinite series of lower bounds, where the members are denoted by $\mu_0, \mu_1, \mu_2, \mu_3, \cdots$, and where μ_0 is identical to alpha and μ_1 is identical to lambda-2 (Guttman, 1945). An important feature of the series is that $\mu_0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \cdots \leq \rho_{XX'}$, with equality when all items or test parts, on which the coefficients are based, are essentially tau-equivalent (see, e.g., Lord & Novick, 1968, p. 90). From empirical results presented by ten Berge and Zegers (1978), no improvement in the first three decimals of μ is to be expected beyond μ_3. Therefore, we will study the sampling behavior of μ_0, μ_1, μ_2 and μ_3 in the present Monte Carlo study.

Subject to rather restrictive conditions, the sampling theory of coefficient alpha has

been derived analytically for dichotomously (Feldt, 1965; Horn, 1971) and polychotomously (Kristof, 1963) scored items. Consequently, we prefer to concentrate on the sample results for alpha in our Monte Carlo study. We realize, however, that a study on the robustness of the distributional properties of alpha derived by Feldt deserves attention in its own right (see Sedere & Feldt, 1977).

Variables of Interest in the Simulation Study

The several estimates of $\rho_{XX'}$, are obviously relevant statistics, and specifically, two dependent variables: First, the average standard deviation of the relevant statistics across a large number of random samples from a known population, which gives an indication of the stability. Second, the average bias with respect to $\rho_{XX'}$, where the average is taken across replications. Denoting statistics by t and their averages by $\bar{t}$, both the average bias and the standard deviation can be combined in the mean squared error:

$$\text{MSE} = \frac{1}{N}\sum_{r=1}^{N}(t_r - \rho_{XX'})^2 =$$

$$= (\bar{t} - \rho_{XX'})^2 + \frac{1}{N}\sum_{r=1}^{N}(t_r - \bar{t})^2, \tag{25}$$

where N denotes the number of random samples. The square root of the mean squared error will also be reported.

The bias of a statistic usually denotes the difference $E(t) - \theta$, where $E(t)$ is the expectation of statistic t across random samples and t is an estimate of the parameter θ. In this study, bias refers to the difference of $E(t)$ and $\rho_{XX'}$, although in a strict statistical sense, t is not always an estimate of $\rho_{XX'}$. Nevertheless, since the reliability estimates were all proposed to approximate $\rho_{XX'}$, we feel that our terminology is justified.

The Simulation Study

The population model. To simulate dichotomous item responses, the two-parameter logistic model by Birnbaum (1968) is used:

$$\pi_i(\xi) = \frac{\exp\left[\alpha_i(\xi - \delta_i)\right]}{1 + \exp\left[\alpha_i(\xi - \delta_i)\right]}, \tag{26}$$

where δ_i and α_i denote the difficulty and discrimination parameters of item i, respectively. Combinations of item parameters must be chosen so that the item characteristic functions do not intersect, which would violate the double monotone condition. Furthermore, we choose the density of ξ to be normal, which is a very common and not unrealistic choice in this kind of research.

The response process. Since our derivations are based on two stochastic processes, we model the simulated response process accordingly. First, values of ξ have to be drawn at random from the normal density. Second, once $\pi_i(\xi_p)$ for person p has been established by means of the two-parameter logistic function, an item score must be drawn at random from his/her propensity distribution. The method of establishing the score of person p on item i is described in detail by van den Wollenberg (1982, p. 126–127).

The design. To evaluate the sampling behavior of the reliability statistics, four independent variables are controlled:

(i) The difficulty parameters. We consider equal distances between the difficulty parameters in each cell of the design. Across cells, the distance varies.

(ii) The discrimination parameters. These are held constant between items and vary

TABLE 2

Standard deviation (SD), bias and square root of the mean squared error (RMSE) of seven reliability estimates.

n=100 ; k=7

		d(δ)=.20							d(δ)=.67						
		Mok-1	Mok-2	MS	alpha	lam-2	mu-2	mu-3	Mok-1	Mok-2	MS	alpha	lam-2	mu-2	mu-3
	SD	1 62	62	62	58	54	54	54	2 81	80	80	77	70	70	70
α=1.00	BIAS	- 5	- 6	- 6	- 6	6	8	8	-13	-16	-11	-22	- 3	1	1
	RMSE	62	62	62	59	55	55	55	82	81	81	80	71	70	70
	SD	3 21	20	19	19	19	18	18	4 40	35	31	33	31	31	30
α=3.00	BIAS	- 1	- 3	- 1	-13	-11	-11	-10	0	-35	1	-87	-53	-45	-44
	RMSE	21	20	20	23	22	21	21	40	49	31	93	61	55	53

Note: the true reliability ($\rho_{XX'}$) equals .589, .534, .882 and .812 in the cells 1, 2, 3 and 4, respectively. The cell number is in the left upper corner of each cell. The entries in the table should be multiplied by 0.001 ; e.g., an entry 69 means 0.069.

TABLE 3

Standard deviation (SD), bias and square root of the mean squared error (RMSE) of seven reliability estimates

		n=100 ; k=15													
		d(δ)=.10							d(δ)=.30						
		Mok-1	Mok-2	MS	alpha	lam-2	mu-2	mu-3	Mok-1	Mok-2	MS	alpha	lam-2	mu-2	mu-3
		[1]							[2]						
	SD	37	37	37	37	35	35	35	39	39	39	38	35	35	35
α=1.00	BIAS	- 1	- 2	- 1	- 3	5	5	5	2	1	1	- 6	6	7	7
	RMSE	37	38	37	37	35	35	35	39	39	39	38	36	36	36
	SD	[3] 9	9	9	9	9	9	9	[4] 16	16	16	17	15	15	15
α=3.00	BIAS	1	1	1	- 5	- 4	- 4	- 4	- 1	- 4	0	-40	-22	-21	-20
	RMSE	9	9	9	10	10	9	9	16	17	16	43	27	25	25

Note: the true reliability ($\rho_{XX'}$) equals .753, .713, .940, and .905 in the cells 9, 10, 11 and 12, respectively. See also the note below Table 2.

across cells in the design.

(iii) The number of subjects. To study the relation between sampling variability and sample size, both relatively small ($n = 100$) and relatively large ($n = 300$) samples are considered.

(iv) The number of items. Since the reliability of a measure is susceptible to the number of items used, tests composed of relatively small ($k = 7$) and large ($k = 15$) numbers of items are considered.

The four independent variables are combined into a $2 \times 2 \times 2 \times 2$ design.

Items within one cell of the design always have different difficulties. This is desirable in the Mokken model, which aims at the ordering of both persons and items on a latent continuum. When two items have the same difficulty in the population, their order in the sample is based purely on coincidence. Furthermore, in practice such items together tend to have a low coefficient of scalability (Jansen, 1982b; also, see Sijtsma and Prins, 1986), which is regarded as undesirable by Mokken.

The Number of Replications

The gain of enlarging the number, N, of replications with respect to accuracy of the sampling distribution decreases with increasing N. We somewhat arbitrarily choose $N = 200$, see also Boomsma (1983, p. 47, 48). As a check on the stability of our results, cross-validations in some cells of the design are carried out.

Results

The standard deviation, the average bias and the square root of the mean squared error of the reliability estimates, are shown in Tables 2 and 3, multiplied by 1000. The entries in the tables are averages across 200 replications of subjects by items data matrices. The tables show results for two levels of item discrimination (α) and two levels of distance between the item difficulties ($d(\delta)$). The number of items varies across tables. The results for different numbers of subjects are summarized in the text.

Distributional Results

Stability results. Given the data of Table 2, on the lines marked "SD", the stability of the three methods discussed in this paper is mostly less than the stability of the classical estimates, though the differences are rather small. The stability of the MS-method is always the same or a little bit better than Mokken's methods, but the differences may be neglected for practical purposes. With respect to the classical estimates, there seems to be a trend towards increasing stability for higher terms in the mathematical series of mu-coefficients. Again, the differences are small.

When $n = 300$, the standard deviation of the estimates has a size of approximately $\frac{1}{2}$ to $\frac{3}{4}$ of the standard deviation in Table 2. Compared to Table 2, when $k = 15$ (Table 3), the standard deviation also decreases considerably. Increasing both the number of subjects and the number of items leads to a standard deviation which has a size of about $\frac{1}{4}$ to $\frac{1}{3}$ of the standard deviation reported in Table 2.

Bias results. The bias of the three methods discussed is mostly smaller than the bias of the classical methods, which is especially true for Method 1 and the MS-method. In Cell 4 of Table 2, the differences are very pronounced: Mokken's Method 1 and the MS-method are almost unbiased while the bias of the classical methods is considerable. The latter perform badly with respect to bias since the items in Cell 4 strongly deviate from essential tau-equivalence. Consequently, the classical parameters are lower bounds to the true reliability.

When $n = 300$, the average biases are of the same magnitude as in Table 2. When $k = 15$ (Table 3) the bias of our three methods is reduced to only a few thousandths, but the bias of the classical estimates in Cell 4 remains considerable. In contrast to the situation of Cell 4 in Table 2, the bias of Method 2 of Mokken may now be considered negligible. Increasing the number of subjects as well does not influence the bias in comparison with the situation where $n = 100$ and $k = 15$ (Table 3).

MSE results. When the items clearly deviate from essential tau-equivalence, the square root of the mean squared error (RMSE) is definitely better for the methods of Mokken and the MS-method. When item discrimination is relatively small, the RMSE of the classical methods is always a little bit better; for relatively large item discrimination, the RMSE of the classical methods is always a little bit worse.

Stability of Results

A cross validation study within several cells of the design was carried out to assess the stability of the results. In general, it turned out that the third decimal place in the Tables 2 and 3 should not be taken very seriously but the conclusions of the study remain unchanged.

Some Additional Results

Since the full results of our study are too numerous to report, we only mention a few interesting secondary results.

First, the correlation across replications between the estimated variance of the raw test score and the reliability estimates is often larger than .90. The correlations are most extreme for the classical estimates, which might result from the sum of all the covariances between test parts figuring both in the numerator and the denominator of the formulas. The three alternative methods do not directly involve these covariances in the numerator, which might explain why they depend less upon the test variance.

Second, the proposed improvements for special cases of the nonparametric methods are often needed in random samples.

In general, when item discrimination is relatively low, estimates of π_{ii} may be below its lower bound. In, for example, Cell 1 of Table 2 this occurred in 43 out of 200 samples.

When the number of subjects is relatively small, the number of samples in which at least two item popularities are equal is relatively large. In, for example, Cells 1 of Tables 2 and 3, this number equals 71 and 192 out of 200, respectively.

The number of times a choice problem (when $\hat{\pi}_i - \hat{\pi}_{i-1} = \hat{\pi}_{i+1} - \hat{\pi}_i$) occurs in Mokken's Method 1, is large especially when the number of subjects is small. In Cell 1 of Table 2, for instance, the choice problem occurs in 67 samples, where the frequency that $\hat{\pi}_{i-1} = \hat{\pi}_i = \hat{\pi}_{i+1}$ is not counted, since this case is counted under the heading of equal proportions correct.

Discussion

For bias, the general picture clearly is in favor of the methods of Mokken and the MS-method. Especially Mokken's Method 1 and the MS-method show negligible bias in all situations under consideration in this study. The classical estimates perform rather poorly when the items strongly deviate from essential tau-equivalence.

We may conclude that the stability of all reliability estimates does not differ very much, although there is a definite trend in favor of the classical estimates. This implies that the classical estimates are more efficient than the estimates discussed in this paper.

Increasing the number of observations always leads to a considerable decrease of the standard deviation, supporting the idea that all estimates may be considered to be consis-

tent. Definite conclusions would call for more variation in the number of subjects.

Theoretically, the double monotone assumption may pose a serious restriction on the usefulness of the estimates presented here. The advice would be then to check the data set on the double monotone property (Mokken 1971, p. 132; Molenaar, 1982a) before estimating the reliability. Finally, Cliff (1983) has pointed out that Mokken's approach to reliability estimation is comparable to approaches advocated by Horst (1953) and Cliff (1984), since the three methods suggest ways for approximating the covariance between replications of parallel items. Also, Mokken's method is the only one that assumes an underlying measurement model, that being the doubly monotone model.

References

Birnbaum, A. (1968). Part V. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality.* Unpublished doctoral dissertation, University of Groningen.

Cliff, N. (1983). Evaluating Guttman scales: Some old and new thoughts. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Cliff, N. (1984). An improved internal consistency reliability estimate. *Journal of Educational Statistics, 9,* 151–161.

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

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30,* 357–370.

Fischer, G. H. (1974). *Einführung in die theorie psychologischer tests* [Introduction to psychological test theory]. Bern: Huber.

Gustafsson, J. E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program* (Internal Rep. No. 63). Institute of Education, University of Goteborg.

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

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.* Princeton: Princeton University Press.

Henning, H. J. (1976). Die Technik der Mokken-Skalenanalyse [The technique of Mokken scale analysis]. *Psychologische Beiträge, 18,* 410–430.

Horn, J. (1971). Integration of concepts of reliability and standard error of measurement. *Educational and Psychological Measurement, 31,* 57–74.

Horst, P. (1953). Correcting the Kuder-Richardson reliability for dispersion of item difficulties. *Psychological Bulletin, 50,* 371–374.

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

Jansen, P. G. W. (1982a). Homogenitätsmessung mit Hilfe des Koeffizienten H von Loevinger: Eine kritische Diskussion [Measuring homogeneity by means of Loevinger's coefficient H: A critical discussion]. *Psychologische Beiträge, 24,* 96–105.

Jansen, P. G. W. (1982b). De onbruikbaarheid van Mokkenschaalanalyse [On the uselessness of Mokken scale analysis]. *Tijdschrift voor Onderwijsresearch, 7,* 11–24.

Jansen, P. G. W. (1983). *Rasch analysis of attitudinal data.* Unpublished doctoral dissertation. Den Haag: Rijks Psychologische Dienst.

Jansen, P. G. W., Roskam, E.E.Ch.I., & Wollenberg, A. L. van den (1982). De Mokkenschaal gewogen [Weighing the Mokken scale]. *Tijdschrift voor Onderwijsresearch, 7,* 31–42.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28,* 221–238.

Lewis, C. (1983). Bayesian inference for latent abilities. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing.* San Francisco: Jossey-Bass.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45,* 507–530.

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233–245.

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

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251–280.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton.

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

Molenaar, I. W. (1982a). Mokken scaling revisited. *Kwantitatieve Methoden, 8*, 145–164.

Molenaar, I. W. (1982b). Een tweede weging van de Mokkenschaal [A second weighing of the Mokken scale]. *Tijdschrift voor Onderwijsresearch, 7*, 172–181.

Molenaar, I. W. (1982c). De beperkte bruikbaarheid van Jansen's kritiek [On the limited usefulness of Jansen's criticisms]. *Tijdschrift voor Onderwijsresearch, 7*, 25–30.

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

Oosterloo, S. (1984). Confidence intervals for test information and relative efficiency. *Statistica Neerlandica, 38*, 37–53.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticisms of classical test theory. *Psychometrika, 42*, 193–198.

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

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. (1984). Useful nonparametric scaling: A reply to Jansen. *Psychologische Beiträge, 26*, 423–437.

Sijtsma, K., & Prins, P. M. (1986). Itemselectie in het Mokken model [Item selection in the Mokken model]. *Tijdschrift voor Onderwijsresearch, 11*, 121–129.

Stokman, F. N., & Schuur, W. H. van (1980). Basic scaling. *Quality and Quantity, 14*, 5–30.

ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika, 43*, 575–579.

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.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology, 32*, 629–658.

Wollenberg, A. L. van den (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123–140.

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

Manuscript received 4/12/85
Final manuscript received 4/9/86