Tijdschrift voor Onderwijsresearch 9 (1984), nr. 5, pp. 257-268.

# Internal Consistency and Reliability in Mokken's Nonparametric Item Response Model

I.W. Molenaar
*Vakgroep Statistiek & Meettheorie F.S.W., Rijksuniversiteit Groningen*

K. Sijtsma
*Vakgroep Arbeids- & Organisatiepsychologie, Vrije Universiteit Amsterdam*

ABSTRACT

The concepts of internal consistency and reliability are not only important in classical test theory, but are equally useful in assessing measurement quality when a latent trait model has been used. They are considered in this paper in the context of Mokken's nonparametric item response model. This leads to a comparison of Loevinger's $H$, Cronbach's $\alpha$ and the – not observable – reliability $\rho_{XX'}$. The assumptions of the model permit an approximation of the reliability in terms of probabilities that correspond to observable events; Mokken's results in this respect are presented and extended.

The concepts of internal consistency and reliability, originally developed in classical test theory, will be considered in this paper within the context of nonparametric item response theory. When a latent trait $\xi$ is measured by $k$ dichotomous items indexed by $i$, this theory is based on the item characteristic function

$$\pi_i(\xi) = P(X_i = 1 \mid \xi) \ , \tag{1}$$

where $X_i$ denotes the observed score of a subject on item $i$, which equals 1 if the response is positive or correct. Contrary to parametric models proposed by e.g. Rasch and Birnbaum, the nonparametric version does not assume a specific functional form for (1) and deals with order relations rather than ratio or interval measurement.

The deterministic Guttman model (1950) can be viewed as an early attempt to order persons and items on a latent trait continuum; here each item has a threshold value $\delta_i$ such that the positive answer is given if and only if $\xi \geqq \delta_i$ (unit stepfunctions in (1)). Each instance in which a negative answer to an easier item and a positive answer to a more difficult item is given, is called an error, and constitutes a deviation from the model.

Mokken (1971) has presented a probabilistic counterpart of Guttman's model, which permits deviations from the perfect model to occur to some limited degree. Besides that, the other assumptions of Mokken's model are:

1. unidimensionality of the latent trait being measured by the items in the test;

Authors' addresses:
I. W. Molenaar, Oude Boteringestraat 23, 9712 GC Groningen
K. Sijtsma, De Boelelaan 1081, 1081 HV Amsterdam

2. local stochastic independence of the answers of a fixed subject to the items;
3. nondecreasing item characteristic functions (monotonicity in $\xi$, across persons);
4. Invariant ordering of the success probabilities on a set of items for all values of $\xi$ (monotonicity across items).

Because of the last two assumptions the model is called doubly monotone (Mokken, 1971, p. 118). An important feature of this model is that item characteristic functions may not intersect, although they may touch. Note that neither the mathematical form of the item characteristic functions, nor any distributional form across subjects of the latent trait being measured is specified.

For recent introductions to the Mokken model see Stokman and Van Schuur (1980) and Mokken and Lewis (1982). The critical comments by Jansen (1982) have led to a discussion in the Tijdschrift voor Onderwijsresearch and in other journals. A key conclusion from this debate, although not essential for the present paper, is that the coefficient $H$ (introduced below in equation (2)), is not a suitable measure for double monotony.

## CONCEPTS OF INTERNAL CONSISTENCY

As argued by Fischer (1974) and others, the quality of measurement is generally improved when it is based on an explicit measurement model. The ideal of expressing measurement quality in one numerical index, however, leads to complications in an item response model: assessing the extent to which the various model assumptions are satisfied may ask for more than one index or test, and moreover does not coincide with assessing the extent to which differences in scores assigned to persons reflect genuine differences in the property being measured.

In many unidimensional latent trait models the magnitude of the overall dependence between items is not essential for the model fit, as long as it is positive; see e.g. Mokken (1971, p. 131), Wood (1978) and Jansen (1983, p. 6-7). More specifically, it can be proved that two doubly monotone items have a nonnegative correlation, and consequently a test complying with the double monotony can have zero reliability as a limiting case. This is not true for the Guttman model, where items correlate maximally given their marginals and the reliability equals one, see below. Fischer (1974, p. 293) and Gustafsson (1977, p. 89) have advocated the use of the test information function for assessing the adequacy of the items for measurement on given levels of the latent trait; see also Lord (1980). This function cannot be used in nonparametric models, because it contains the latent subject and item parameters. For the doubly monotone model, however, Mokken (1971, p. 142, 148) has presented means for assessing global dependence of a set of items. In the sequel we will investigate some properties of these means and present some new developments.

Two kinds of internal consistency will be distinguished in this section, and measures for both kinds will be compared.

In the foregoing section we have already mentioned Guttman-homogeneity as a form of internal consistency of a data set. This form is the more restrictive one of the two to be mentioned, in demanding that all correlations between items be maximal given their marginals. Coefficient $H$ (Loevinger, 1948; Mokken, 1971; Van Naerssen, 1972) is a measure of this kind of internal consistency. It can be written as a linear function of the sum across all possible item pairs of the proportions of errors:

$$H = 1 - \frac{\sum\sum_{i<j} \pi_{ij}}{\sum\sum_{i<j} \pi_i (1 - \pi_j)} \quad . \tag{2}$$

In this formula $\pi_{ij}$ is the proportion of subjects answering the more difficult item $i$ correctly and the easier item $j$ incorrectly. The item difficulties are denoted by $\pi_i$ and $\pi_j$ ($\pi_i < \pi_j$), where $\pi_i$ denotes the proportion answering item $i$ correctly. It will be assumed that $0 < \pi_i < 1$ for all $i$.

When the data comply with the Guttman model $H$ equals one, meaning that the data contain no errors. When the answering behaviour of subjects is random, $H$ will equal zero. This will become clear if one rewrites $H$ as

$$H = \frac{\sum\sum_{i<j} \Phi_{ij}}{\sum\sum_{i<j} \Phi_{ij}\,(\max)} \quad , \tag{3}$$

where $\Phi_{ij}$ denotes the well-known product-moment correlation between dichotomous items. The fact that a zero value of $H$ indicates random behaviour and thus absence of internal consistency in the data, can be considered as an advantage of $H$ over other coefficients of Guttman-homogeneity (see White & Saltz, 1957 and Mokken, 1971 Ch. 2). Another advantage of $H$ is that its asymptotic sampling theory is known (Mokken, 1971, p. 157-169) as well as its exact distribution per item pair (Molenaar, 1982, p. 154-155). Jansen (1982, 1983) has put forward that $H$ is not a good index of double monotony, and that an item set which satisfies the Rasch model may well have a low value of $H$.

The other concept of internal consistency is the classical one, based on the magnitude of the correlation between items. One may argue that positivity of such correlations is sufficient, because in classical test theory sense a long test of lowly correlating items can be as good as a short test of highly correlating items. Practical considerations such as time limitations in taking and scoring the test, as well as the limited availability of suitable items, restrict the usefulness of this argument. On the other hand, the so-called attenuation paradox shows that extremely high item correlations may decrease the validity of the test, see e.g. Lord & Novick (1968, p. 344).

Deferring a more general discussion of the not directly observable reliability to a later section, the present section deals with the well known coefficient $\alpha$ (Cronbach, 1951), called *KR20* for our case of dichotomous items, as a measure of internal consistency:

$$\alpha = \frac{k}{k-1} \cdot \frac{2\sum\sum_{i<j} \sigma_{ij}}{\sigma^2(X)} \quad , \tag{4}$$

where $k$ is the number of items in the test, $\sigma_{ij}$ is the covariance between items $i$ and $j$ and $\sigma^2(X)$ is the variance of the test scores. Coefficient $\alpha$ can also be written as a curvilinear function of the sum of the proportions of errors over all possible item pairs. Using the fact that $\sigma^2(X)$ can be written as the variance of a linear combination of item scores and furthermore the property of 2x2 – contingency tables that $\pi_{ij} = \pi_i - \pi_{ij}^{-}$, where $\pi_{ij}$ denotes the proportion answering both items $i$ and $j$ correctly,

$$\alpha = \frac{k}{k-1} \cdot \frac{2 \sum\sum_{i<j} \pi_i (1-\pi_j) - 2 \sum\sum_{i<j} \pi_{i\bar{j}}}{2 \sum\sum_{i<j} \pi_i (1-\pi_j) + \sum_i \pi_i (1-\pi_i) - 2 \sum\sum_{i<j} \pi_{i\bar{j}}}, \tag{5}$$

with $\pi_i < \pi_j$. For given values of the item difficulties this formula only contains positive constants except for the error proportions: a simpler formulation would be

$$\alpha = \frac{k}{k-1} \cdot \frac{a - 2 \sum\sum_{i<j} \pi_{ij}}{a + b - 2 \sum\sum_{i<j} \pi_{i\bar{j}}}, \tag{6}$$

where $a > 0$ and $b > 0$.

Coefficient $\alpha$ can only reach its maximum of 1 in the case of dichotomous item scores if all item difficulties are equal. It is in this case only that the items can correlate perfectly. It should be noted, however, that for practical purposes the effect of unequal item difficulties on the value of $\alpha$ does not seem to be very dramatic (Cronbach, 1951, p. 327). Like $H$, coefficient $\alpha$ equals zero in the case of random behaviour of the respondents. Both $H$ and $\alpha$ do not have fixed minima; their negative magnitudes are restricted by the fact that $\sigma^2(X) > 0$. But this is no problem, because negative values of $H$ and $\alpha$ would point to the fact that one or more items are negatively associated with other items, which is undesirable in psychological tests.

## COMPARISON OF $H$ AND $\alpha$

Although $H$ and $\alpha$ should not be interchanged in research, because they reflect different kinds of internal consistency, several authors have commented on their apparent similarities (e.g. Cronbach, 1951; Horst, 1953; Guilford, 1954; Cliff, 1977; Terwilliger and Lele, 1979; Cudeck, 1980; Raju, 1982). The situations in which both coefficients are zero or one have already been discussed. For a numerical comparison in intermediate situations, one must assume certain aspects to be constant and others to vary freely: Some of the authors quoted above have made different and not always plausible decisions in this respect.

Cronbach (1951, p. 326) and Carroll (1961, p. 323) study the behaviour of $H$ for an item pair under constant tetrachoric correlation. Jansen (1982, 1983) varies the popularity of items with parallel logistic item characteristic curves. From his study it can be concluded that for almost coinciding curves $H$ will be low but $\alpha$ may be high. Some authors assume a constant variance or a constant distribution of total scores, others assume constant item popularities. It should be noted that such different assumptions often lead to different conclusions.

In this section we will compare $H$ and $\alpha$ when holding constant a set of mutually different item difficulties. It follows from (2) and (6) that $H$ and $\alpha$ can then be expressed as functions of the proportion of errors in score patterns. From Figure 1 and 2 it becomes clear that $H$ is larger than $\alpha$ only if the number of items is very small and the variance of the item difficulties is large (Figure 1). But if we take 6 (Figure 1) or 5 items (Figure 2), $\alpha$ is mostly larger than $H$. This becomes even more pronounced when the number of items increases. Only for extremely low proportions of errors will $H$ be larger than $\alpha$. For the special case when all item difficulties are equal $H$ never exceeds $\alpha$. We may conclude then that for realistic test lengths, say 10 or more items, $H$ will amost always be smaller than $\alpha$.

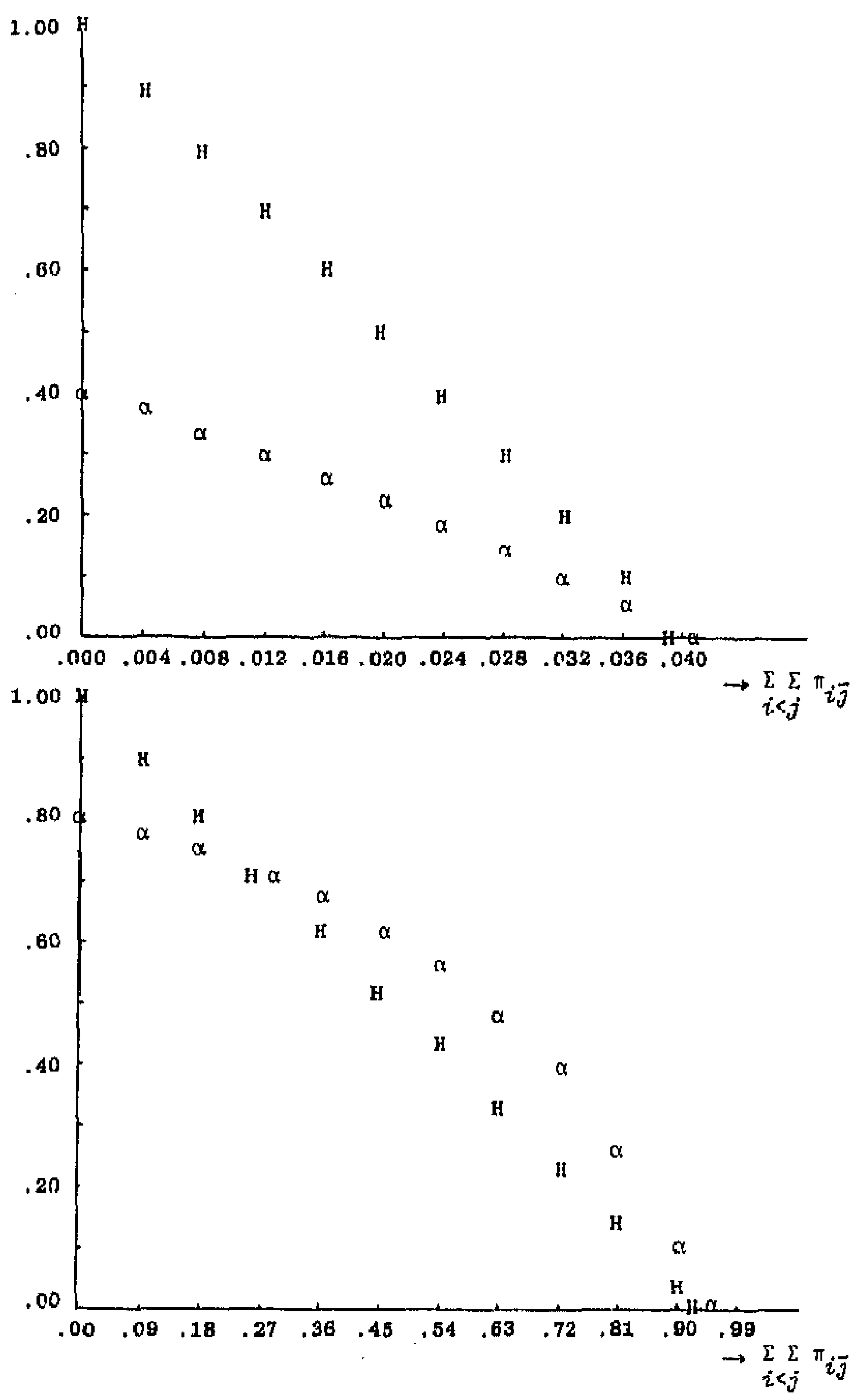


Figure 1: $H$ and $\alpha$ plotted as functions of the proportion of errors.
Upper graph: $k = 2$ ($\pi = .2; .8$); lower graph: $k = 6$ ($\pi = .1; .2; .3; .7; .8; .9$).

In Cudeck's (1980) Monte Carlo study, $H$ and $\alpha$ are compared by means of artificial datasets that were generated according to the Birnbaum model (1968). Some of his results are summarized in his Figure 4, where both coefficients are plotted against the mean item discrimination in the simulated test. It turns out that his results, which are necessarily inexact, are essentially comparable to ours. This becomes clear especially when it is realized that the mean item discrimination of a test will increase if the number of errors (our Figures 1 and 2) decreases.

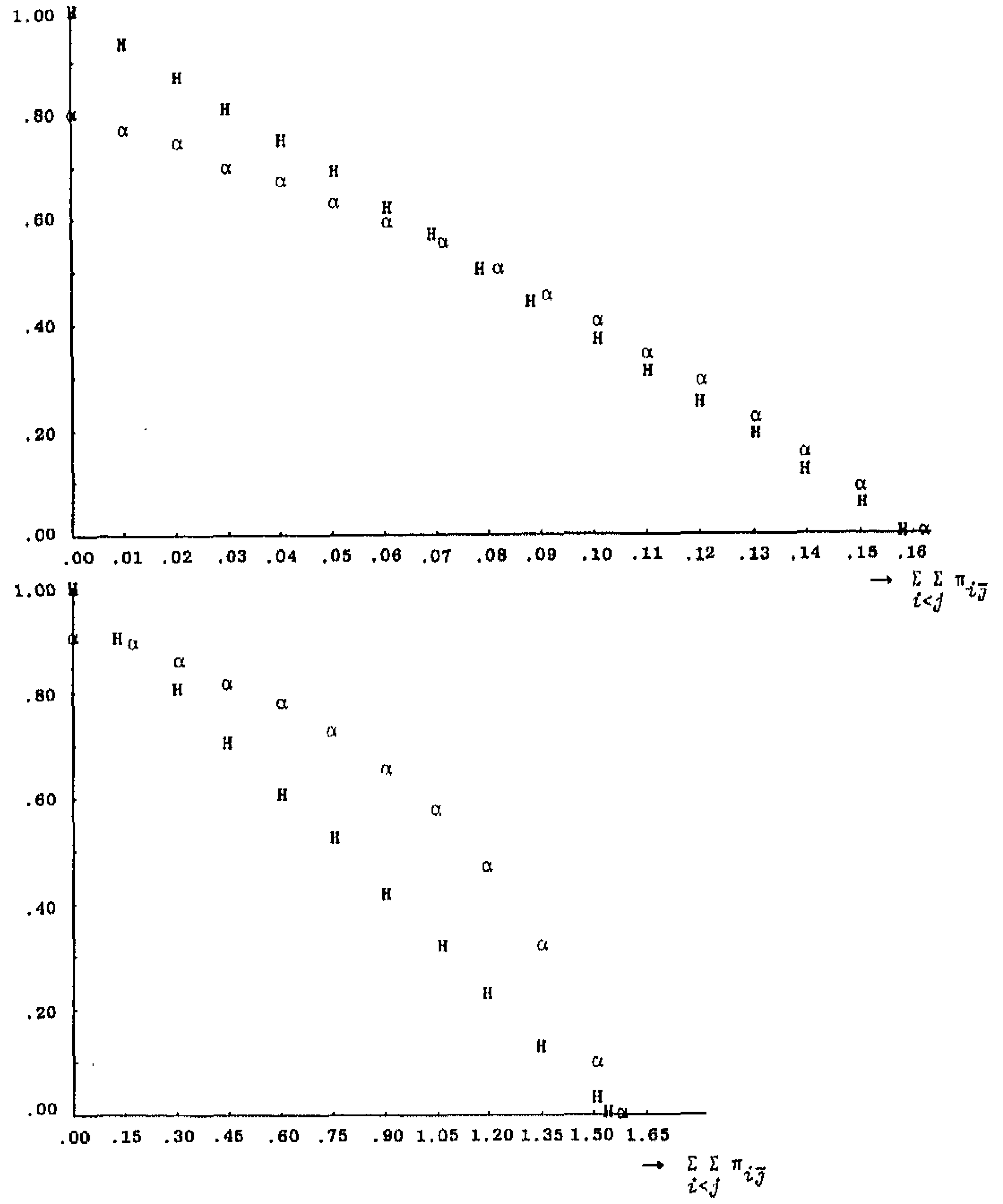


Figure 2: $H$ and $\alpha$ plotted as functions of the proportion of errors.
Upper graph: $k = 2$ ($\pi = .4; .6$); lower graph: $k = 5$ ($\pi = .3; .4; .5; .6; .7$).

## RELIABILITY AND PARALLELLISM

In classical test theory the observed score $X$ is written as the sum of the true score and the error. In item response theory, this means that the true score for a person with latent trait value $\xi$ becomes $E(X \mid \xi) = \sum_i \pi_i(\xi)$. If $E_\xi$ and $\sigma_\xi^2$ denote the mean and variance with respect to the distribution of $\xi$ across persons, one may use for the population of persons the well known variance decomposition

$$\sigma^2(X) = \sigma_\xi^2[E(X \mid \xi)] + E_\xi[\sigma^2(X \mid \xi)] , \tag{7}$$

which separates the systematic variation (between persons, due to the latent trait) from the average error variation (within persons). In practical applications of classical test theory it is usually assumed that $\sigma^2 (X \mid \xi)$ is independent of $\xi$. This assumption, however, is not realistic for extreme values of $\xi$ because of floor and ceiling effects. Proceeding without it in item response theory, it remains informative to use the reliability coefficient

$$\rho_{XX'} = \frac{\sigma^2_\xi [E(X \mid \xi)]}{\sigma^2 (X)} = 1 - \frac{E_\xi [\sigma^2 (X \mid \xi)]}{\sigma^2 (X)} \ , \tag{8}$$

for assessing the proportion of the total variance that is explained by the latent trait.

Neither theory has direct solutions for the problem that latent variables such as $\xi$ or $E(X \mid \xi)$ are not observable. In both theories one may resort to split-half or test-retest estimates of the reliability, which have their well known difficulties. A more popular alternative is the use of coefficient $\alpha$: it is well known that this is a lower bound to the reliability. Other estimation methods, especially suited to nonparametric item response models, will be discussed in the next section.

Coefficient $\alpha$ is equal to the reliability if and only if all items are essentially tau-equivalent (Lord & Novick, 1968, p. 90), i.e. when constants $a_{ij}$ exist such that for each item pair and for all $\xi$

$$E(X_i \mid \xi) = E(X_j \mid \xi) + a_{ij} \ . \tag{9}$$

For our dichotomous items this means, in the notation (1),

$$\pi_i (\xi) = \pi_j (\xi) + a_{ij} \ , \tag{10}$$

and thus the item characteristic curves must be 'vertically parallel', meaning pictorially that they lie above each other with a constant vertical distance between them (Figure 3). They can be translated into each other by moving the curves along the vertical axis. As illustrated in Figure 3, this requires asymptotes strictly between zero and one, and will nearly always be even less plausible than the already demanding assumption of 'horizontal parallellism' $\pi_i (\xi) = \pi_j (\xi + b_{ij})$ postulated in the Rasch model. Horizontal parallellism pictorially means that the item characteristic curves lie next to each other and can be translated into each other by simply moving the curves along the horizontal axis. If nearly all persons have their latent trait value between a and b in Figure 3, one may expect more violation of (10) for the steep curves. The difference between reliability and $\alpha$ also depends on other factors, however.

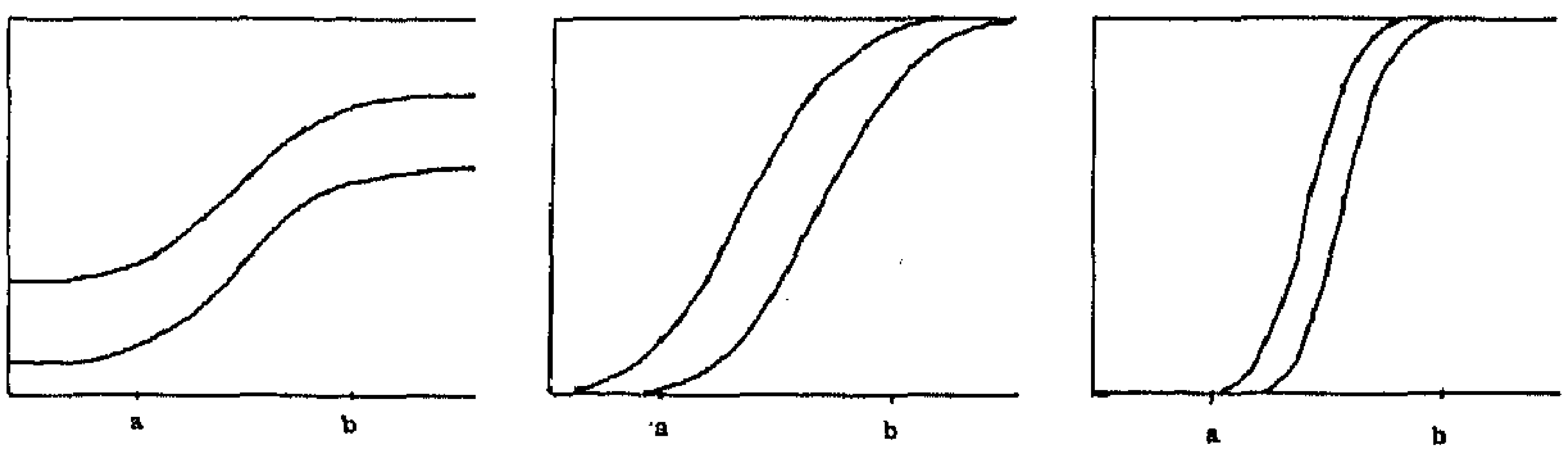


Figure 3: Vertically parallel (left) and horizontally parallel (middle and right) item characteristic curves. If the latent trait values lie mostly in the interval (a, b), differences in vertical distance are moderate for moderate slopes, but dramatic for steep curves.

A more extreme but also relevant case is a test consisting of parallel (Lord & Novick, 1968, p. 50) rather than essentially tau-equivalent items: the item curves now completely coincide because the true scores are equal. The correlation coefficient $\rho_{ij} = \rho(X_i, X_j)$ is the same for each item pair, and substitution of the equal popularities and equal correlations leads to $H = \rho_{ij}$ and $\rho_{XX'} = \alpha = k\rho_{ij}/(1 + (k-1)\rho_{ij})$, a well known psychometric formula showing clearly that in this extreme case $H$ is constant but $\alpha$ and $\rho_{XX'}$ increase with test length.

In the general case one may expect that $H$ is high for steep curves, and given the steepness mostly for curves widely spaced compared to the variation in the latent trait (much variation in popularities). The test length has little or no systematic influence on $H$ because an increase in the number of errors is compensated by an increase of the score variance. Coefficient $\alpha$ is high for steep and nearly coinciding curves; it comes close to the reliability for nearly coinciding curves. Both the reliability and its lower bound $\alpha$ rise with increasing test length although the rise levels off for long tests. The foregoing statements are simplifications because of the complicated interplay of the various factors.

In this paper we have restricted our attention to coefficient alpha because it is the measure of internal consistency most widely in use. Other coefficients have e.g. been proposed by Ten Berge and Zegers (1978) who present an infinite family of coefficients of which $\alpha$ and Guttman's $\lambda_2$ are members. An approach to estimate the greatest lower bound to the reliability of a test is presented by Ten Berge, Snijders and Zegers (1981).

## RELIABILITY ESTIMATES IN THE MOKKEN MODEL

As explained by Mokken (1971, p. 132, 146), the assumption of double monotony permits an estimation method for $\rho_{XX'}$ not based on replication or internal consistency. For any given $\xi$ the score $X$ is the sum of $k$ locally independent item scores with variances $\pi_i(\xi)[1-\pi_i(\xi)]$, thus

$$\sigma^2(X \mid \xi) = \sum_i \pi_i(\xi)[1-\pi_i(\xi)] = \sum_i [\pi_i(\xi) - \pi_i^2(\xi)] \ . \tag{11}$$

Taking expectations with respect to $\xi$ one obtains $E_\xi[\sigma^2(X \mid \xi)]$, which can be inserted into (8):

$$\rho_{XX'} = 1 - \sum_i (\pi_i - \pi_{ii})/\sigma^2(X) \ , \tag{12}$$

where

$$\pi_i = \int \pi_i(\xi)\, dG(\xi) \quad \text{and} \quad \pi_{ii} = \int \pi_i^2(\xi)\, dG(\xi) \ , \tag{13}$$

with $G$ denoting the cumulative distribution function of $\xi$. This in itself has not solved our problem, because $\pi_{ii}$ is not observable.

The observed sample popularity of item $i$ is an unbiased and consistent estimate of $\pi_i$. The same holds for the fraction of persons answering positively to items $i$ and $j$ ($j \neq i$) as an estimate of

$$\pi_{ij} = \int \pi_i(\xi)\, \pi_j(\xi)\, dG(\xi) \ . \tag{14}$$

Estimation of $\pi_{ii}$ now proceeds by replacing one of the two factors $\pi_i(\xi)$ in the integrand in (13) by an approximation which is a linear function of one or two $\pi_j(\xi)$ for $j \neq i$: if

$$\pi_i(\xi) \approx a + b\pi_j(\xi) + c\pi_h(\xi) \quad (j \neq i,\ h \neq i)\ , \tag{15}$$

then

$$\pi_{ii} \approx \int [a + b\pi_j(\xi) + c\pi_h(\xi)]\ \pi_i(\xi)\, dG(\xi) = a\pi_i + b\pi_{ij} + c\pi_{ih}\ , \tag{16}$$

which contains only observable probabilities.

It remains to find a suitable linear combination. As the double monotony implies

$$\pi_{i-1}(\xi) \leqq \pi_i(\xi) \leqq \pi_{i+1}(\xi) \quad \text{for all } \xi\ , \tag{17}$$

two simple candidates are

$$\pi_i(\xi) \approx \pi_{i-1}(\xi)\, \pi_i / \pi_{i-1} \quad \text{thus} \quad \pi_{ii} \approx \pi_{i-1,i}\, \pi_i / \pi_{i-1}\ , \tag{18}$$

and

$$\pi_i(\xi) \approx \pi_{i+1}(\xi)\, \pi_i / \pi_{i+1} \quad \text{thus} \quad \pi_{ii} \approx \pi_{i,i+1}\, \pi_i / \pi_{i+1}\ . \tag{19}$$

Mokken (1971, p. 147) recommends as 'method 1' to choose (18) if $\pi_i$ is closer to $\pi_{i-1}$ than to $\pi_{i+1}$ and (19) otherwise. For $i = 1$ and $i =$ k there is only one neighbour and therefore no choice is left.

A second candidate linear combination is the interpolation

$$\pi_i(\xi) \approx \pi_{i-1}(\xi) + [\pi_{i+1}(\xi) - \pi_{i-1}(\xi)]\ (\pi_i - \pi_{i-1}) / (\pi_{i+1} - \pi_{i-1}), \tag{20}$$

and thus

$$\pi_{ii} \approx \pi_{i-1,i} + (\pi_{i,i+1} - \pi_{i-1,i})\ (\pi_i - \pi_{i-1}) / (\pi_{i+1} - \pi_{i-1})\ . \tag{21}$$

The use of (21), completed by (18) for $i = k$ and (19) for $i = 1$, is called 'method 2' by Mokken. Note that (19) is found from (21) by replacing $\pi_{i-1}(\xi)$ by the function that is identically zero, by which $\pi_{i-1}$ and $\pi_{i-1,i}$ are zero too, and similarly (18) results from (20) by replacing $\pi_{i+1}(\xi)$ by the zero function.

This exposition of Mokken's reliability estimation methods leads to the following questions:

A) do the methods change when the direction of the scale is reversed?

B) are they accurate (both in absolute sense and compared to $\alpha$)?

C) if not, how can they be improved?

Ad A) (reversal). Reversing the direction of the scale means e.g. that a high score now must be interpreted as political conservatism instead of liberalism. It is clear that a Mokken scale remains a Mokken scale when negative and positive answers are interchanged; this simply means that zero and one change roles and each $\pi$ must be replaced by $1 - \pi$. A little algebra shows that the interpolation result (21) remains unaltered, but (18) and (19) now become

$$\pi_{ii} \approx \pi_{i-1,i}\ (1 - \pi_i)/(1 - \pi_{i-1}) + \pi_i\ (\pi_i - \pi_{i-1})/(1 - \pi_{i-1})\ , \tag{22}$$

$$\pi_{ii} \approx \pi_{i,i+1}\ (1 - \pi_i)/(1 - \pi_{i+1}) - \pi_i\ (\pi_{i+1} - \pi_i)/(1 - \pi_{i+1})\ , \tag{23}$$

which can also be obtained from (21) by using the function identically equal to one (rather than zero as above) as a replacement of $\pi_{i-1}$ ($\xi$) or $\pi_{i+1}$ ($\xi$). It follows that substituting (22) and (23), or (18) and (19) in the reliability formula (12) will lead to different reliability estimates.

Ad B) (accuracy). In a simulation study two-parameter logistic curves for $\pi_i$ ($\xi$) and a standard normal distribution for $\xi$ have been used to obtain all relevant probabilities (including $\pi_{ii}$ and the exact $\rho_{XX'}$) by numerical integration. These are population values as no sampling procedure was involved. It results that Mokken's method 2 nearly always underestimates the true reliability, with an error roughly half the error of coefficient $\alpha$. Method 1 produces some positive and some negative errors, of about the same magnitude as in method 2. Errors when using $\alpha$ typically range between –.01 and –.06 for item curves with slopes equal to the standard deviation of $\xi$ (in our study 1), but can be as large as –.16 for slopes three times as large. In another study, the item parameters of ten items were chosen in such a way as to replicate the empirical $\pi$-values and $H_{ij}$-values of 10 items from an attitude scale towards school ($n$ = 2000). The true reliability in the simulated case equals .82 and is underestimated by only .02 for $\alpha$ and .01 for either Mokken method.

Ad C) (improvement). A detailed study of the errors shows that underestimation and overestimation per individual $\pi_{ii}$ mostly cancel each other when the average of the four quantities (18), (19), (22) and (23) is used, with obviously (18) and (22) only for $i = k$ and (19) and (23) only for $i = 1$. The absolute error of this average was in all examples between 1/3 and 1/5 of the error of $\alpha$ when that was small, and at most 1/10 of the error of $\alpha$ when that was large (say .08 or more). For hand calculations most of the gain can be achieved already by averaging (18) and (19) for $2 \leqq i \leqq k - 1$, proceeding as above for $i = k$ and $i = 1$.

Although space does not permit a full motivation of these proposals, Figure 4 may be of some help. For the three items in the left hand part, the linear combination $^1/_2\pi_{i-1}$ ($\xi$) + $^1/_2\pi_{i+1}$ ($\xi$) of type (21) overestimates $\pi_i$ ($\xi$) below its median and underestimates its upper half. As it is multiplied by $\pi_i$ ($\xi$) $dG$ ($\xi$) before being integrated, the underestimation dominates the overestimation in the final approximation of $\pi_{ii}$.

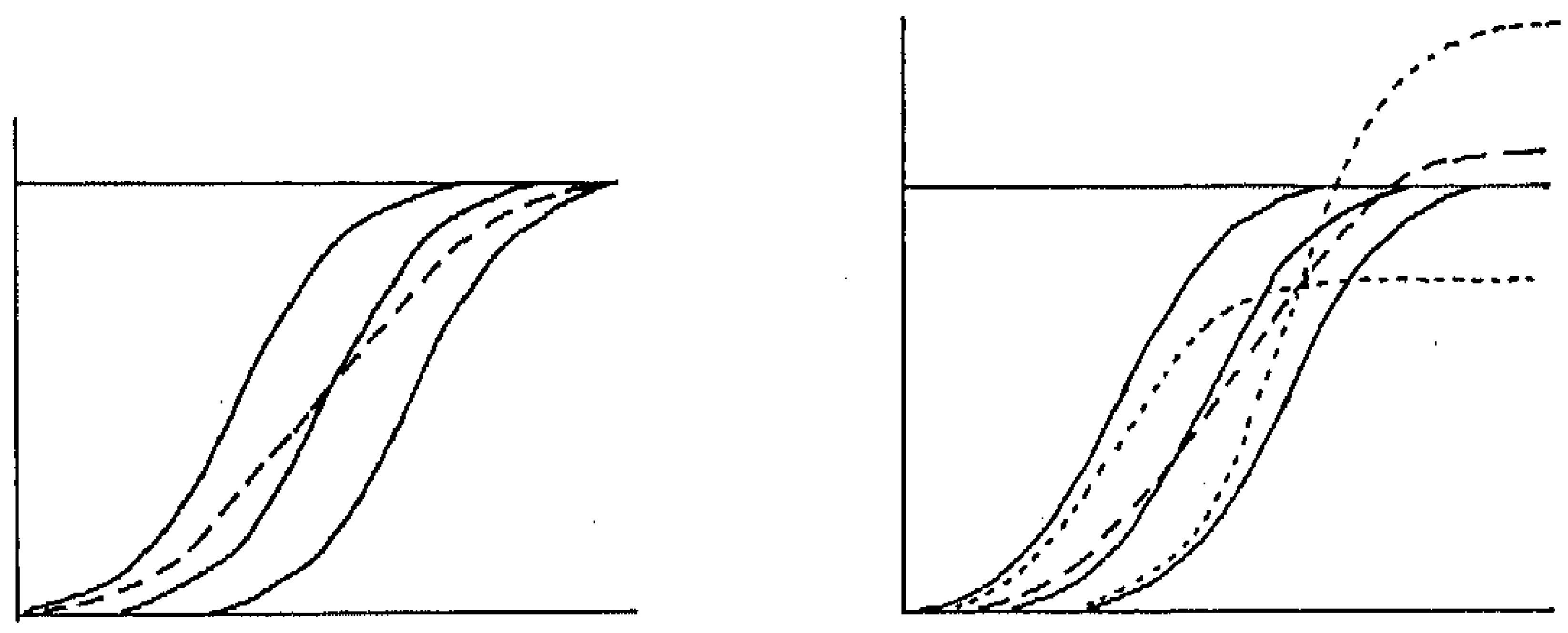

Figure 4: Three curves with $\pi_{i+1}$ = .64, $\pi_i$ = .50 and $\pi_{i-1}$ = .36, illustrating approximation (20) (left, striped curve), approximations (18) and (19) (right, dotted curves) and the average of (18) and (19) (right, striped curve) to $\pi_i$ ($\xi$).

The right half of the figure illustrates (18) and (19) for the same three items. Their average, denoted by the striped curve, is close to $\pi_i$ ($\xi$) in the region which by the factor $\pi_i$ ($\xi$) $dG$ ($\xi$) contributes most to the integral. Multiplicative corrections such as (18) lead to an upper asymptote unequal to one. The numerical results show that this is not a serious threat to the accuracy, unless the scale contains items with popularities very close to zero or one (which are not informative anyway).

By a similar figure, not reproduced here, one could motivate why the average of (18) and (22) works well for $i = k$ or the average of (19) and (23) for $i = 1$.

## DISCUSSION

The functions of, and the relations between, Loevinger's $H$, Cronbach's $\alpha$ and the reliability $\rho_{XX'}$ were explored in the context of Mokken's nonparametric item response theory. It was argued that this model permits its own estimation methods for $\rho_{XX'}$, which were extended and improved. Work is now in progress on two related issues: the extension of the reliability estimation to the case of more than two ordered response categories announced in Molenaar (1982) and the investigation of the small sample properties of the estimates, of which only the population values were compared in the present paper.

The authors agree with Jansen (1983) that the Rasch model, when applicable, offers many advantages not shared by the Mokken model. The latter has partially different and often less demanding assumptions, however, and this has motivated the study of its measures for internal consistency and reliability in the present paper.

## REFERENCES

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

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

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

Carroll, J.B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347-372.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika, 42,* 375-401.

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

Cudeck, R. (1980). A comparative study of indices for internal consistency. *Journal of Educational Measurement, 17,* 117-130.

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

Guilford, J.P. (1954). *Psychometric methods.* New York: McGraw-Hill.

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

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. *Measurement and prediction.* Princeton: Princeton University Press.

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

Jansen, P.G.W. (1982). De onbruikbaarheid van Mokkenschaalanalyse. *Tijdschrift voor Onderwijsresearch, 7,* 11-24.

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

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

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. (1982). Mokken scaling revisited. *Kwantitatieve Methoden, 3,* (8), 145-164.

Naerssen, R.F. van. (1972). Eenvoudige formules voor de optimale spreiding van item-p-waarden. *Nederlands Tijdschrift voor de Psychologie, 27,* 123-133.

Raju, N.S. (1982). On tests of homogeneity and maximum KR-20. *Educational and Psychological Measurement, 42,* 145-152.

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

Terwilliger, J.S., & Lele, K.T. (1979). Some relationships among internal consistency, reproducibility and homogeneity. *Journal of Educational Measurement, 16,* 101-108.

White, B.W., & Saltz, E. (1957). Measurement of reproducibility. *Psychological Bulletin, 54,* 81-99.

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