KM 8(1982)
pag 145-164



# MOKKEN SCALING REVISITED

Ivo W. Molenaar *)

Summary

Mokken (1971) has proposed a nonparametric latent trait model for unidimensional scaling of dichotomous items. The present paper reports on current research aimed at extension and refinement of this model. The case of three or more ordered answer categories per item is worked out in section 2. Some problems in Mokken's search procedure for tables with low expected frequencies are discussed in section 3. Statistical procedures for a detailed investigation of the monotonicity assumptions of the model are proposed in section 4, including their relation to goodness of fit tests for the Rasch model. The effects of conditioning on the observed item popularities are the subject of section 5, which is followed by a discussion (section 6). Because the research project is not yet finished, some of the material presented has a preliminary character.

---

*) Vakgroep Statistiek en Meettheorie FSW, Oude Boteringestr. 23, 9712 GC Groningen. Thanks are due to Boomsma, Lewis, Van Schuur and Verbeek for comments on an earlier version and to ms. S. Kroonenberg for programming help.

## 1. Introduction

Let a random sample of n persons answer k dichotomous items ($X_i$=1 or 0 if a person's answer to item i is positive or negative). Let all items measure one continuously distributed latent trait $\theta$, and let the item curves $P(X_i=1|\theta)$ be non-decreasing and non-intersecting. For this situation Mokken (1971) has proposed a nonparametric model, based only on local independence and monotonicity, allowing a probabilistic extension of the Guttman scale in which all item curves would be unit step functions. We refer to Mokken (1971), Stokman & Van Schuur (1980), and Mokken & Lewis (1982) for a complete description of the model, which includes a variety of estimation, test and search procedures embedded in a computer program (STAP User's Manual, 1980).

Within the assumptions the item curves may vary between two extremes: unit step functions in the deterministic Guttman model, and horizontal lines in the null model of stochastically independent answers (not only locally for persons with equal $\theta$, but globally for all persons). When allowing some exceptions from the former, the procedures seek to safeguard against the latter: totally unrelated items do not measure the same trait and should be eliminated.

With capitals denoting random variables, the four answer patterns for two items have the following probabilities and frequencies in the general model:

probabilities

| | $X_j=0$ | $X_j=1$ | |
|---|---|---|---|
| $X_i=0$ | $\pi_{ij}$ | $\pi_{\bar{i}j}$ | $1-\pi_i$ |
| $X_i=1$ | $\pi_{i\bar{j}}$ | $\pi_{ij}$ | $\pi_i$ |
| | $1-\pi_j$ | $\pi_j$ | 1 |

frequencies

| | $X_j=0$ | $X_j=1$ | |
|---|---|---|---|
| $X_i=0$ | $A_{ij}$ | $B_{ij}$ | $n-N_i$ |
| $X_i=1$ | $F_{ij}$ | $N_{ij}$ | $N_i$ |
| | $n-N_j$ | $N_j$ | $n$ |

The reason for the special notation $F_{ij}$ and $N_{ij}$ will be explained below. For $\pi_i < \pi_j$, Mokken uses Loevinger's H as the population scalability coefficient for the item pair:

$$H_{ij} \stackrel{def}{=} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i(1-\pi_j)} = \frac{phi}{phi_{max}} \qquad \text{for } \pi_i < \pi_j. \tag{1.1}$$

In this notation the null case means $\pi_{ij}=\pi_i\pi_j$ and $H_{ij}=0$. The distribution of $N_{ij}$ given $N_i=n_i$ and $N_j=n_j$ would then be hypergeometric $(n,n_i,n_j)$, and Mokken refers the normalized statistic

$$\Delta^*_{ij} = \frac{N_{ij}-n_in_j/n}{\{n_in_j(n-n_i)(n-n_j)n^{-2}(n-1)^{-1}\}^{\frac{1}{2}}} \tag{1.2}$$

to standard normal tables, which is an asymptotic ($n \to \infty$) conditional test of the null hypothesis.

At the other extreme, the Guttman case means $\pi_{i\bar{j}}=0$, thus $\pi_{ij}=\pi_i$ and $H_{ij}=1$. There are then no persons in the "error cell", scoring 1 on item i and 0 on the easier item j. The number of persons $F_{ij}$ actually observed in this error cell is thus a count of the number of violations of the Guttman property (notation F for "fault"). Such faults may be summed for a fixed person across item pairs for detection of outlying persons with unusual answer patterns (cf. Molenaar (1982a) for the Mokken, or Wright & Stone (1979) for the Rasch model). In the Mokken procedures, $F_{ij}$ is summed across j for assessing the scalability of item i, and across all pairs $i < j$ for assessing the overall scalability; in both cases a suitable normalization is added that need not concern us here.

Inserting $\pi_{ij}=\pi_i-\pi_{i\bar{j}}$ into (1.1) one obtains

$$H_{ij} = 1 - \frac{\pi_{i\bar{j}}}{\pi_i(1-\pi_j)} \quad , \tag{1.3}$$

showing the close relation of $H_{ij}$ to the ratio of actual error probability and null hypothesis error probability.

Inserting sample fractions, this leads to the estimate

$$\hat{H}_{ij} = 1-F_{ij}/e_{ij} \tag{1.4}$$

where

$$e_{ij} = n_i(n-n_j)/n \qquad (n_i < n_j) \tag{1.5}$$

is the conditional null expectation of the error frequency.

To this very brief recapitulation of the existing publications on the Mokken model it should be added that its qualifications as a measurement model have been debated in the Tijdschrift voor Onderwijsresearch, see Molenaar (1982c) for references. The present paper, partly elicited by this debate, contains four different amendments and extensions to the Mokken scaling theory and procedures.

Section 2 points out that the limitation to dichotomous items is an undesirable restriction in the frequent applications to data with more than two answer categories. For all cases where such categories are ordered, it contains an extended model which allows the application of the Mokken principles without previous dichotomization of the answers.

Section 3 signals that even for moderate or large sample size n, there may well be some item pairs for which at least one null-expected frequency in the 2x2 table is low. The estimate $\hat{H}_{ij}$ (1.4) is then unreliable and the distribution of $\Delta^*_{ij}$ (1.2) is far from normal. Amendments based on the exact distribution or the Poisson approximation are proposed.

Section 4 reports on current research on the double monotony (non-decreasing and non-intersecting item curves) which is an essential assumption that can only be indirectly checked in the data matrix.

Section 5 deals with a possible problem stemming from the application of the standard tools of normalization and conditioning to our scaling situation. If the statistical inference aims at describing the behavior in repeated samples of size n from a population with success probabilities $\pi_i$, $\pi_j$ and jointly $\pi_{ij}$ for the two items i and j, the 2x2 table should be described by a quadrinomial distribution of the four cell frequencies. Under the null hypothesis $N_{ij}$ then has a binomial $(n, \pi_i \pi_j)$ distribution rather than hypergeometric $(n, n_i, n_j)$. Under any hypothesis, null or alternative, the conditioning on the marginals obscures that $N_i > N_j$ may be observed even when $\pi_i < \pi_j$, especially when $\pi_j - \pi_i$ is small. In such a sample the investigator would use the wrong error cell $X_i=0$, $X_j=1$ rather than $X_i=1$, $X_j=0$. This may lead to wrong inferences in the null case, and to overestimation of $H_{ij}$ in alternative cases. Our addition of $\pi_i < \pi_j$ to (1.1) and $n_i < n_j$ to (1.5) is not always correct; in a nonnegligible number of cases the sample ordering of items may contradict the population ordering and bias the inference.

## 2. Three or more ordered categories

Suppose each of k items has three ordered answer categories, such as {disagree, neutral, agree} for an attitude item or {wrong, partially right, correct} for an achievement item (an extension to more than three categories, and to a different number of categories per item, will be dealt with at the end of this section). One could run a Mokken, Guttman or Rasch analysis after a dichotomization which assigns the middle category to one of the remaining categories. This means loss of information, fewer possible total score values for the measurement of persons, and possibly different results for different dichotomization decisions; examples are given in Molenaar (1982a ). In Likert scaling, one would assign values 0,1,2 to the three categories. Gifi (1981) on the contrary assigns "best" category values according to a complicated least squares loss function, for which no inferential statistical procedures are known. The aim of this section is to explore how it can be ascertained whether the Likert scaling values are compatible with the assumptions of an extended Mokken model, and thus lead to a probabilistic ordering of items and persons on one latent continuum.

It has long been recognized that each three-category item may be viewed as two consecutive "item steps": as a first step the subject might ascertain whether (s)he has enough of the latent trait to score at least a one, rather than zero. If and only if this is the case, the second item step is tried, in which it is established whether the trait value allows for a two rather than a one on the original item. This division into item steps is completely described for the Rasch model by Masters (1982), who also compares it to approaches based on category boundaries, e.g. Samejima (1969), Andrich (1978, 1979).

Our aim is to call a set of multicategory items a Mokken scale if and only if their dichotomous item steps form a Mokken scale. This demands an exact definition of such steps. Let the score X on a three category item be 0,1,2 with probabilities a,b,c respectively (a+b+c=1). The first item step score $S_1$ is 0 for X=0 and 1 for $X \geq 1$. For the

second item step, one may either choose a conditional scoring $S_{2*}$ which is treated as a missing value for persons failing the first step, or a cumulative scoring $S_2$ which is 0 for $X \leq 1$ and 1 for X=2. Masters (1982) presumably aims at a Rasch model for $S_1$ and $S_{2*}$, in which $P(S_{2*}=1)=b/b+c$ can be either smaller or larger than $P(S_1=1)=b+c$. Here we shall use the cumulative scoring, in which the important relation $X=S_1+S_2$ holds for any subject. Note that there is a functional dependence between both item steps: $S_1=0$ implies $S_2=0$ and $S_2=1$ implies $S_1=1$.

If one would treat k items with three ordered categories as 2k separate dichotomous item steps, a straightforward application of the Mokken procedures to these 2k "new items" is invalidated by this dependence. For any pair of two item steps from the same original item, the 2x2 table has a logically impossible error cell, and the null assumption of independence is meaningless. In the matrix of $H_{ij}$ values for the 2k item steps, the entries for two steps from the same item are automatically equal to one. This would artificially inflate the $H_i$ values per item step and the H value for the total scale.

A better solution is illustrated in the following fictitious 3x3 table of two three category items (table 2.1). The easiest item step $X_j \geq 1$ is passed by 70 percent, next $X_i \geq 1$ by 60 percent, next $X_i \geq 2$ by 50 percent and finally $X_j \geq 2$ by 30 percent. If all item steps possessed the Guttman property, subjects with a growing value of the latent trait would consecutively be in one of the five"perfect" cells joined by line segments. Each subject in one of the four remaining cells, marked by stars, has passed a certain item step but failed an easier one, and cannot be assigned without error to the joint representation of item steps and score patterns which is displayed in table 2.1b. The 20 persons with $X_i=0$ and $X_j=2$, for example, have failed the easy step $X_i \geq 1$, passed by 60%, but succeeded at the step $X_j \geq 2$ passed by 30%. With dichotomous items there would be three perfect patterns and only one error cell. Here, however, we extend the estimation of the Guttman conformity of the two items i and j, by summing across the error cells. Let $F_{ij}$ and $e_{ij}$ denote the observed and null-expected total frequency of respondents in the four error cells. Then the proposed estimate of scalability per item pair is calculated as:

$$\hat{H}_{ij} = 1 - \frac{F_{ij}}{e_{ij}} = 1 - \frac{10+10+20+10}{30+150+120+30} = 1 - \frac{50}{330} = 0.85 \qquad (2.1)$$

Table 2.1a. Joint distribution of two three-category items, with expected numbers under independence in brackets, and cumulative percentages passing the step. Consecutive perfect patterns are joined by line segments and error cells are marked by a star.

| | $X_j=0$ | $X_j=1$ | $X_j=2$ | total | cumul. percentage passed |
|---|---|---|---|---|---|
| $X_i=0$ | 280(120)—— | 100(160) | *20(120) | 400 | 100% |
| $X_i=1$ | *10( 30) | 80( 40) | *10( 30) | 100 | 60% |
| $X_i=2$ | *10(150) | 220(200)— | 270(150) | 500 | 50% |
| total | 300 | 400 | 300 | 1000 | |
| Cumulative percentage passed | 100% | 70% | 30% | | |

Table 2.1b. Joint representation of item steps and perfect Guttman patterns; comparison of observed frequency of such patterns to expected frequency under both the Guttman and the null model.

| pattern $(X_i,X_j)$ | $X_j \geq 1$ (0,0) | $X_i \geq 1$ (0,1) | (1,1) | $X_i \geq 2$ (2,1) | $X_j \geq 2$ (2,2) | total perfect |
|---|---|---|---|---|---|---|
| observed | 280 | 100 | 80 | 220 | 270 | 950 |
| Guttman exp. | 300 | 100 | 100 | 200 | 300 | 1000 |
| null exp. | 120 | 160 | 40 | 200 | 150 | 670 |

Next, the item scalability $H_i$ of item i and the overall scalability H in the pool of k items are estimated by summing across item j and across all pairs $i < j$, as before.

The following instruction can be generally applied for a set of k items in which item i has $m_i+1$ ordered categories ($m_i \geq 1$):

A) Divide each item into steps $X_i \geq m$ $(m=1,2,\ldots,m_i)$

B) For each item pair (i,j) put the $m_i+m_j$ item steps in descending order of popularity and join the corresponding cells in their cross tabulation like in table 2.1

C) If $F_{ij}$ and $e_{ij}$ denote the observed and null-expected frequency respectively summed across the remaining cells of the table, calculate $\hat{H}_{ij}$ by (2.1).

This instruction leads to the generalized concept of a Mokken scale for multicategory items, in which persons and item steps are ordered on the same continuum. The position of a person on this ordinal scale is the total number of item steps passed, and this is just the Likert score obtained by scoring $X_i=0,1,\ldots,m_i$ for the categories of item i. A person with total score $\Sigma X_i=Y$ should in most cases have passed the Y easiest item steps and failed the $\Sigma m_i-Y$ other item steps. The total number of violations F of this property should be lower than expected under independence of the items, as is expressed by $H = 1-F/e$ in which e is the expected number summed across cells and across item pairs.

The publications (e.g. Mokken, 1971, section 5.2) and the computer programs (STAP User's Manual, 1980) offer for dichotomous items the following six possibilities:

1) the evaluation of a set of items as one scale;
2) the construction of a scale from a given pool of items (search procedure);
3) multiple scaling (the construction of a number of scales from a given pool of items);
4) the extension of a given scale by means of a larger pool of items;
5) the investigation of the double monotony or holomorphism of a set of items;
6) the computation of reliability coefficients.

For procedures 1) to 4) the multicategory case now presents no additional problems, except for the asymptotic null distribution of $\hat{H}_{ij}$ used in the search procedure, discussed on the next page. Procedure 5

will be discussed in section 4, in which its generalization to more than two categories will be straightforward . Mokken's estimates of the classical reliability coefficient (only a by-product of his model) will only be applicable in the multicategory case after some modification not pursued in this paper.

Mokken derives the asymptotic null distribution of $\hat{H}_{ij}$ for the dichotomous case from the hypergeometric distribution of $F_{ij}$ or $N_{ij}$, see section 1. Now that $F_{ij}$ becomes a sum across the $m_i m_j$ error cells of a $(m_i+1)x(m_j+1)$ table with $m_i+m_j+1$ perfect patterns, its null distribution should be established, as it plays a role in the procedures 1)to 4). The following tentative solutions are proposed:

1) Asymptotically, the multinomial distribution of all entries becomes multinormal with the restriction of total sum equal to n. Conditioning on the marginals means imposing some more linear constraints on the entries, and the conditional mean and variance of $F_{ij}$ might be calculated using the estimated correlations between cell entries resulting from the constraints; this would give a normal approximation for $F_{ij}$.
2) The very fast computer program FISHER by Kroonenberg & Verbeek (1981) generates all r*c tables with given marginals and orders them by their Pearson $\chi^2$ value; it has been modified into ordering by $H_{ij}$, leading to the exact distribution of this quantity given the marginals. A first try-out of the modified program shows that cases with a large n and/or more than three categories per item may produce more than $10^7$ tables. A favorable 3x3 case with n=679 had $5*10^5$ tables and took 7 CP seconds on the CDC Cyber 170/760. It may be desirable to first divide all observed frequencies by 2 or even 5, rounding upward for the error cells; significance in this smaller table is sufficient but not necessary for significance in the original table.
3) As was remarked in section 1, it is dubious whether the conditioning on the marginals offers the best way for predicting the behavior of a Mokken scale in a future sample of respondents. In the unconditional distribution, the notation $\phi_{im}=P(X_i=m)$ enables us to state that $F_{ij}$ is

binomial $(n,\pi)$ with, for the case of table 2.1,

$$\phi = \phi_{i0}\phi_{j2} + \phi_{i1}\phi_{j0} + \phi_{i1}\phi_{j2} + \phi_{i2}\phi_{j0} .$$

Note that the location of the error cells in the table varies with the item marginals. Inserting the estimated item marginals would lead to at least a rough binomial test for the observed fraction $F_{ij}/n$ in the combined error cells. See section 5 for more discussion. With one or more of these solutions implemented the ordered categories extension of the Mokken model seems to be ready for empirical try-outs and for implementation in the standard computer program.

## 3. Small expected frequencies

In this section we return to the original Mokken model for two categories. In the search procedure, a scale is stepwise built up from a given pool of items; at each stage a new item i added to an already formed subscale should have a positive $\hat{H}_{ij}$ with all items j in the subscale and a significantly positive $\hat{H}_i$ with respect to this subscale. It will be demonstrated that both requirements are possibly too restrictive when cases with low expected frequencies are involved.

Suppose that among several items administered to a sample of n=100 subjects there are two items with population value $H_{ij}=0.5$ for which $n_i=8$ and $n_j=90$ subjects give the positive answer. The estimated item popularities of 0.08 and 0.90 are somewhat extreme, but still realistic. As outlined in section 1, the null expected frequency in the error cell is $e_{ij}=n_i(n-n_j)/n=0.8$. From the observed frequency $F_{ij}$ in this cell one calculates the estimate $\hat{H}_{ij}=1-F_{ij}/e_{ij}$; note that it equals 1 for $F_{ij}=0$, but is negative for any $F_{ij} \geq 1$. If $H_{ij}$ were zero then $F_{ij}$ would have a hypergeometric distribution in which one obtains $P(F_{ij}=0)=0.42$, but of course it is no problem that a sample estimate $\hat{H}_{ij}$ then is negative for 58 percent of the samples. But if our true $H_{ij}$ is 0.5, the expected value of $F_{ij}$ under this alternative equals $(1-H_{ij})e_{ij}=0.4$, and the exact distribution of $F_{ij}$ for the given marginals now is extended

hypergeometric, see Harkness (1965) or Molenaar (1982 b sec.4), with estimated odds ratio 0.4*82.4/(7.6*9.6), and $P(F_{ij}=0)$ is then calculated to be 0.66. Thus although the population value $H_{ij}$ is 0.5, there is still a probability of 0.34 that its sample estimate is negative. In about one third of the samples the researcher would infer that the two items are negatively associated and thus may never be admitted to the same Mokken scale, even if their H with the remaining items were high.

The problem is indeed caused by small $e_{ij}$. An analysis for $H_{ij}$=0.5, $n_i$=19, $n_j$=80 leads to $e_{ij}$=3.8 and a probability of 0.10 for $F_{ij} \geq 4$, which is now the condition for a negative estimate of $H_{ij}$. Similarly for $n_i$=26, $n_j$=70 one gets $e_{ij}$=7.8 and a probability of only 0.02 for $\hat{H}_{ij} < 0$ with $H_{ij}$=0.5. Figure 3.1 gives detailed graphs of this probability for $0 < e_{ij} < 4$ and four representative values of $H_{ij}$. It is based on the Poisson approximation to the exact distribution of $F_{ij}$, which is very accurate for small expected values but may have an error of up to 0.02 for $e_{ij}$ close to 4. It should not be used for larger values.

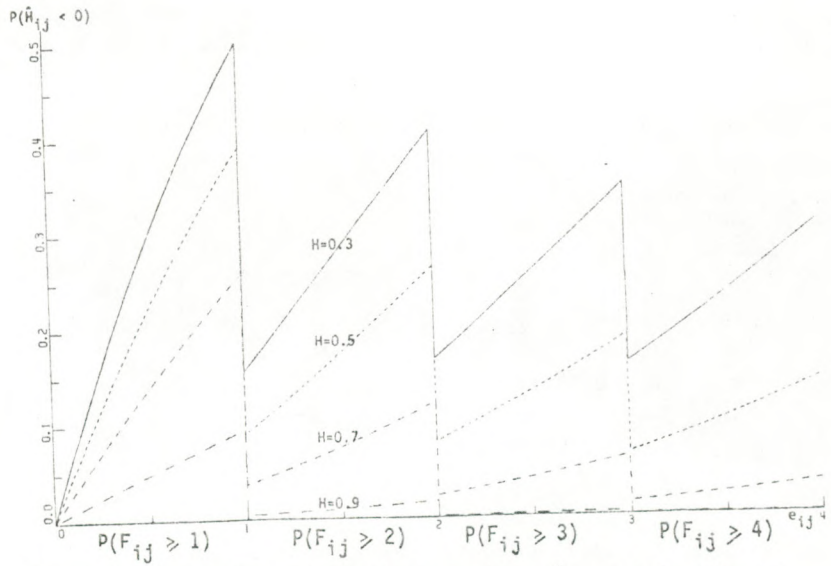


Figure 3.1 Probability of a negative estimate $\hat{H}_{ij}$ as a function of the null expectation $e_{ij}$ for four alternative values of the population value $H_{ij}$ = H (Poisson approximation).

We conclude that though a Mokken scale demands a positive population value $H_{ij}$, requirement of positivity for its sample estimate is too restrictive for item pairs with low $e_{ij}$, say $e_{ij} \leq 10$. The user of the Mokken computer programs should be given the opportunity, when such item pairs exist, to inspect the exact or approximate probabilities under meaningful alternatives like $H_{ij}=0.3$, 0.5 or 0.7, and to optionally override the strict requirement that two items with a negative estimate $\hat{H}_{ij}$ may never appear in the same scale. An additional advantage of requiring $e_{ij} > 10$ is that $\hat{H}_{ij}=1-F_{ij}/e_{ij}$ then decreases by steps of less than 0.1 when $F_{ij}=0,1,2,....$

We next turn to the related requirement that $H_{ij}$ should be significantly positive for the first item pair admitted, and $H_i$ should be significantly positive with respect to the items already in the scale for adding item i to it. For protection of the total error rate of these tests, the individual significance levels take into account the total number of tests executed during the search process, as detailed by Mokken (1971, p.196-197).

For an example of the influence of a low expected frequency we turn to the data in table 3.1 for nine items measuring pupils' self judgment on their academic achievement in high school, see Molenaar (1982a, 1983).

Table 3.1 Observed frequencies $N_i$ (italics, on the diagonal) and $N_{ij}$ (lower triangle) for nine achievement items answered by 679 pupils; the upper triangle gives $\hat{H}_{ij}$ for each pair.

| | V21 | V16 | V35 | V25 | V3 | V12 | V36 | V23 | V13 |
|---|---|---|---|---|---|---|---|---|---|
| V21 | *48* | 0.25 | 0.46 | 0.31 | 0.34 | 0.37 | 0.61 | 0.88 | 1.00 |
| V16 | 15 | *57* | 0.16 | 0.28 | 0.17 | 0.12 | 0.38 | 0.71 | 0.15 |
| V35 | 29 | 22 | *184* | 0.22 | 0.37 | 0.31 | 0.45 | 0.29 | 0.65 |
| V25 | 24 | 27 | 80 | *186* | 0.12 | 0.17 | 0.35 | 0.50 | 0.91 |
| V3 | 26 | 24 | 103 | 72 | *206* | 0.52 | 0.16 | 0.21 | 0.76 |
| V12 | 27 | 22 | 96 | 79 | 137 | *207* | 0.27 | 0.27 | 0.92 |
| V36 | 37 | 36 | 124 | 114 | 103 | 117 | *277* | 0.49 | 0.53 |
| V23 | 46 | 51 | 137 | 152 | 147 | 152 | 226 | *433* | 0.70 |
| V13 | 48 | 54 | 180 | 185 | 203 | 206 | 269 | 425 | *637* |

The Mokken search procedure is supposed to start by selecting the two items with the highest $\hat{H}_{ij}$, provided that it differs significantly from zero. In this example the procedure skips the item pair V13, V21 with $\hat{H}_{ij}$=1.00 in favor of V12, V13 with $\hat{H}_{ij}$=0.92, because by formula (1.1) the former has $\Delta^*_{ij}$=1.84 and the latter $\Delta^*_{ij}$=4.08. If a level $\alpha$ = 0.05 is used for the whole procedure, this first test is carried out at level $\alpha/\binom{9}{2}$ = .0014, for which the normal deviate equals 2.99. The 2x2 table for V13 and V21 is given in Table 3.2 with the left tail of the exact hypergeometric (679, 42, 48) distribution for the number $F_{ij}$ in the error cell and its normal approximation from $\Delta^*_{ij}$.

This shows not only the inadequacy of the normal approximation (notwithstanding the sample size of 679, due to extremely skewed distributions), but it also exhibits that for the given marginals even the optimal result of 0 in the error cell does not meet significance level of $\alpha/\binom{K}{2}$ = 0.0014 in the exact test.

Table 3.2 The 2x2 table for V13 and V21 (expected numbers under independence in brackets) and left tail of exact and approximate null distribution of $F_{ij}$.

| | V13=0 | V13=1 | total |
|---|---|---|---|
| V21=0 | 42(39) | 589(592) | 631 |
| V21=1 | 0( 3) | 48( 45) | 48 |
| | 42 | 637 | 679 |

| $F_{ij}$ | exact. prob. | normal approx. | $\hat{H}_{ij}$ |
|---|---|---|---|
| 0 | 0.042 | 0.033 | 1.00 |
| 1 | 0.142 | 0.077 | 0.66 |
| 2 | 0.232 | 0.163 | 0.33 |
| 3 | 0.240 | 0.234 | -0.01 |
| $\geq$ 4 | 0.345 | 0.493 | $\leq$ -0.35 |

Stated otherwise, the null expectation of the error cell is only 48*42/679 = 2.97. This makes it impossible to situate a rejection region of size .0014 in the left tail, due to the discreteness and skewness of the exact hypergeometric distribution. Such extreme cases

can be detected from the $\hat{H}_{ij}$ and $\Delta^*_{ij}$ matrices optionally printed in Mokken's computer program; by using such an item pair as a forced start set for the search procedure one can override their omission based on a lack of significance "that they cannot really help". Moreover, if all items really form a scale, then item pairs left out for lack of significance at the first step will enter at later steps of the search procedure; the significance requirement is then put on the item scalability coefficient $H_i$ of a new item with regard to all items already in the scale, and this will nearly always involve larger expected frequencies and less skewed distributions. Nevertheless, the computer program might well print out a special warning in cases where the item with a higher H is skipped in favor of a lower but significant value, and allow the user the option to override this rule. For both effects of small expected frequencies, this section has produced some tentative solutions, of which the side effects ask for more study.

## 4. The assumption of double monotony

Let $\pi_i(\theta) = P(X_i=1|\theta)$ be the trace line of the dichotomous item i. The Mokken model assumes that each $\pi_i(\theta)$ is a nondecreasing function of $\theta$, and that the trace lines of different items do not intersect:
(a) if $\theta_0 < \theta_1$ then $\pi_i(\theta_0) \leqslant \pi_i(\theta_1)$ for each i;
(b) $\pi_i(\theta_0) < \pi_j(\theta_0)$ for one value $\theta_0$ implies $\pi_i(\theta) \leqslant \pi_j(\theta)$ for all $\theta$.
We have chosen this weak definition of double monotony, including equality, because we want to include perfect Guttman items as a limiting case.

Note that (b) implies that the ordering of the items is specifically objective: in any group of persons, item i is not easier than item j. This property of the nonparametric Mokken model corresponds to the stronger requirement in the parametric Rasch model that the ratio of the item difficulties must be invariant across person groups. Thus empirical verification of the monotony assumptions can proceed among similar lines as the goodness-of-fit investigation in the Rasch model. From the more

detailed discussion in Molenaar (1982c) we mention that the "splitter-item procedure" compares the order of all the other items in the two subgroups which have one particular item correct and wrong respectively, and is equivalent to the inspection of the P- and $P_0$-matrix already proposed by Mokken. An external splitting criterion like sex or age may also be used. Finally, in analogy with the Andersen test for the Rasch model, one may split into score groups using the internal criterion of total number of positive answers as an indicator of the latent trait . Score groups also lead to a check of the property (a) of monotonicity per trace line. If the number of items is small this check can be improved by deleting the item actually examined in the formation of score groups, just as item-rest-correlation is more adequate than item-test correlation.

In all these instances, presented in more detail in Molenaar (1982c, 1983), an observed order is compared to a predicted order. The observed order, however, is based on the doubly stochastic scaling model: firstly a sample of subjects is observed from a population for which scalability is desired, and secondly each subject v with latent trait value $\theta_v$ plays for each of the k items an independent chance game, with success probability $\pi_i(\theta_v)$ for the i-th item. The question arises whether discrepancies between observed and predicted order can be ascribed to these two chance mechanisms. Molenaar (1982c) shows how McNemar and Fisher exact tests can be applied to the various 2x2 tables arising from such order comparisons.

The use of formal significance tests for such decisions poses a number of well known problems to which no fully satisfactory solution seems to exist. We mention first the choice of the null hypotheses: a liberal view would only reject the model when a violation cannot be ascribed to chance, a rigid view would require that each order relation is beyond doubt in the predicted direction. Next there is a combination of tests problem: each grouping of subjects leads to a multitude of dependent tests per item and per pair of groups, and moreover one may choose an almost infinite number of groupings. Unless we have a perfect Guttman scale, the worst of these groupings will surely exhibit violations. Research on the implication of these dilemmas for the Mokken monotonicity

investigation is now in progress. In our view it should ideally be based on a decision theoretic loss model, in which the damage of using a scale in which an assumption is violated should be compared to the damage of incorrectly rejecting a scale which has no substantial violations in the population.

## 5. Conditioning on the marginals: the wrong error cell

Choices between a conditional and an unconditional procedure for estimation and hypothesis testing have led to many debates among statisticians, in which different answers are given to questions like "what do you really want to know" and "how well does it work". For the 2x2 table, and more generally the $r \times c$ table, the received view seems to be that conditioning on the observed marginals is advisable when an analysis of (in)dependence is desired; see e.g. Lehmann (1959, sec. 4.6) and Bishop, Fienberg & Holland (1975, Ch.2).

In the statistical analysis of the Mokken model this trend has been followed. But now an additional problem arises in passing from $N_{ij}$ to $\hat{H}_{ij}$, as was announced at the end of section 1. In the present section we shall assume throughout that $\pi_i < \pi_j$, and use the shorthand notation $\bar{i}$ for $X_i=0$, writing e.g. $\pi_{\bar{i}j}$ for $P(X_i=0, X_j=1)$. The population scalability value $H_{ij}$ is given by (1.2). By straightforward algebra it can be rewritten as

$$H_{ij} = (\pi_{ij}\pi_{\bar{i}\bar{j}} - \pi_{\bar{i}j}\pi_{i\bar{j}})/\{\pi_i(1 - \pi_j)\} , \qquad (5.1)$$

in which one recognizes the phi/phi$_{max}$ property. An obvious sample estimate is obtained by inserting sample fractions: multiplying above and below by $n^2$ it becomes

$$\hat{K}_{ij} = \frac{N_{ij}N_{\bar{i}\bar{j}} - N_{\bar{i}j}N_{i\bar{j}}}{N_i(n - N_j)/n} = 1 - \frac{F_{ij}}{N_i(n - N_j)/n} \qquad (5.2)$$

Note that $\hat{H}_{ij}$ as defined in section 1 is equal to $\hat{K}_{ij}$ as long as $N_i \leq N_j$, but when $N_i > N_j$ the denominator of $\hat{H}_{ij}$ becomes $N_j(n - N_i)/n$, and $N_{\bar{i}j}$ is

used instead of $F_{ij} = N_{ij}$, because then the cell $(i,j)$ is incorrectly taken to be the error cell. In those cases $\hat{H}_{ij} = c\hat{K}_{ij}$ with $c = N_i(n-N_j)/\{N_j(n-N_i)\} > 1$. The distinction between $\pi_i < \pi_j$ and $N_i < N_j$ seems to have been ignored by previous authors, although it may have been a reason to allow an ordering enforced by the user in the program MOKKEN TEST (STAP User's Manual, p. SCS 24).

The probability of a wrong observed order is not always small. Computer calculations show that it equals 0.18 for a realistic example like $n = 100$, $\pi_i = 0.45$, $\pi_j = 0.50$, $H_{ij} = 0.30$, It decreases with increasing n, with increasing spacing $\pi_j - \pi_i$, with $(\pi_i + \pi_j)/2$ further from 0.5 and with increasing positive association between items i and j. Further computations will be reported in a later paper.

Under the null model of independence, each conditional distribution of $\hat{K}_{ij}$ has mean zero, because then $E\, F_{ij} = N_i(n-N_j)/n$. This unbiased estimation of the true $H_{ij}$ (which is zero) holds also for $\hat{H}_{ij}$, even when the wrong error cell is used. As the significance test of $H_{ij} = 0$ is purely based on the expression (1.2) for $\Delta^*_{ij}$, which is symmetric in i and j, this conditional test remains valid in the samples in which $N_i > N_j$ was observed.

The difference between $\hat{K}_{ij}$ based on the true error cell and $\hat{H}_{ij}$ based on the sample error cell is more important under the alternative $H_{ij} \neq 0$. Here the absolute value of $H_{ij}$ is overestimated by a factor $c = N_i(N-N_j)/\{N_j(n-N_i)\}$ for all samples with $N_i > N_j$. Whether this is serious depends both on the frequency with which wrong orders are observed, and on the size of c. In the standard procedures 1) to 4) listed in section 2 such estimates are used in checking whether observed H values exceed the lower bound supplied by the user (default 0.3) and in selecting the best item for extending a scale . They also occur in the non-null analysis leading to confidence bounds. In such applications, then, the user should be cautioned that for item pairs with almost the same popularity a repetition in a next sample might exhibit a reversal of the item order, and that some error patterns may therefore not have been noticed as such. On the other hand for such an item pair $N_i$ will probably be close

to $N_{\cdot j}$, thus c will not be too far from 1, and $N_{ij}$ will not differ much from $N_{\hat{i}j}$. A further study on the results for the non-null case is now in progress.

## 6. Discussion

The possibility of using more than two ordered answer categories seems promising. A word of warning should perhaps be given that it will increase the likelihood of bad fit, of low answer frequencies, and of almost equal popularities. It may occur, moreover, that some steps of an item fit very well into the model whereas others do not (e.g. a neutral category may cause problems). Nevertheless, the five items on pupils' relations to classmates described in Molenaar (1982a) provide an example of a good three-category Mokken scale.

The analysis of small frequency cases in section 3 has led to some proposals for an improved computer program. Ideally each analysis could use the normal approximations when appropriate and pass to the exact discrete distributions when necessary.

The proposed tests for the monotonicity requirements are another example of a step forward that is by no means the final step. Not only should a satisfactory combination procedure be added, but also it might be desirable to consider an extension of the model: for item pairs with almost the same popularity one could refrain from predicting their order and from counting inversions of the answer pattern within such pairs as errors. The resulting model for specifically objective partial order would do away with most of the "wrong order problem" discussed in section 5.

In many respects, this progress report contains more problems than answers. It should be viewed as just a step on the road to a thorough statistical investigation of Mokken's nonparametric latent trait model, which has been successfully applied in a number of scaling problems, such as Mokken (1971) on political efficacy, Clason (1977) on sex role differencing, Stokman (1977) on voting by United Nations delegates and

De Vries-Griever c.s. (1982) on sleep quality.

Our paper has stressed, more than Mokken himself, the probabilistic Guttman aspect and double monotony of a Mokken scale, which lead to a falsifiable measurement model. Examples of detailing H into an analysis of errors per pattern, as presented in Molenaar (1982a, 1983) are in line with this view of a Mokken scale as an instrument for a specifically objective joint ordering of persons and items (or item steps).

References

Andrich, D. (1978) A rating scale formulation for ordered response categories, Psychometrika, 43, 561-573.

Andrich, D. (1979) A model for contingency tables having an ordered response classification. Biometrics, 35, 403-415.

Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975) Discrete Multivariate Analysis. MIT Press, Cambridge (Mass.).

Clason, C.E. (1977) Beroepsarbeid door gehuwde vrouwen, dissertatie Rijksuniversiteit Groningen.

De Vries-Griever, A.H.G. , de Vries G.M. & Meyman, T.F. (1982) De Nederlandse Rijksloods. Deel II: Slaap. Groningen, Subfaculteit Psychologie, Vakgroep Arbeids- en Organisatie Psychologie.

Gifi, A. (1981) Nonlinear multivariate analysis. Afd. Data-theorie FSW, Universiteit van Leiden.

Harkness, W.L. (1965) Properties of the extended hypergeometric distribution. Annals of Mathematical Statistics, 36, 938-945.

Kroonenberg, P. & Verbeek, A. (1981) User's Guide to Fisher, University of Utrecht, Department of Mathematics, Preprint nr. 188.

Kroonenberg, P. & Verbeek, A. (1981) FISHER, een programma voor de analyse van r * c tabellen bij kleine steekproeven. Kwantitatieve Methoden, 2, no.2, 66-86.

Lehmann, E.L. (1959) Testing Statistical Hypotheses. Wiley,New York.

Lewis, C. (1981) Estimating abilities: inference for random variables. Kwantitatieve Methoden, 2, no.2, 17-24.

Masters, G.N. (1982) A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

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

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

Molenaar, I.W. (1982a) Steviger meten met een latent model. Jaarboek 1982, Nederlandse Vereniging voor Marktonderzoekers.

Molenaar, I.W. (1982b) Some improved diagnostics for failure of the Rasch model. Psychometrika, in press.

Molenaar, I.W. (1982c) Een tweede weging van de Mokkenschaal. Tijdschrift voor onderwijsresearch, 7, 172-181.

Molenaar, I.W. (1983) Rasch, Mokken en Schoolbeleving. To appear.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement, no. 17.

STAP User's Manual, Vol.4. (1980)(Stochastic Cumulative Scaling, Mokken Scale, Mokken test). Technisch Centrum FSW, Universiteit van Amsterdam.

Stokman, F.N. (1977) Roll calls and sponsorship. A methodological analysis of third world group formation in the United Nations. Sijthoff, Leiden.

Stokman, F.N. & Van Schuur, W.H. (1980) Basic Scaling. Quality & Quantity, 14, no.1, 5-30.

Wright, B.D. & Stone, M.H., Best test design, Rasch measurement, Chicago: Mesa Press, 5835 Kinbarg Avenue, 1979.