

# A FORTRAN 77 program for a nonparametric item response model: The Mokken scale analysis

JOHANNES KINGMA

*University of Utah, Salt Lake City, Utah*

and

TERRY TAERUM

*University of Alberta, Edmonton, Alberta, Canada*

A nonparametric item response theory model—the Mokken scale analysis (a stochastic elaboration of the deterministic Guttman scale)—and a computer program that performs this analysis are described. Three procedures of scaling are distinguished: a search procedure, an evaluation of the whole set of items, and an extension of an existing scale. All procedures provide a coefficient of scalability for all items that meet the criteria of the Mokken model and an item coefficient of scalability for every item. Four different types of reliability coefficient are computed both for the entire set of items and for the scalable items. A test of robustness of the found scale can be performed to analyze whether the scale is invariant across different subgroups or samples. This robustness test serves as a goodness of fit test for the established scale. The program is written in FORTRAN 77. Two versions are available, an SPSS-X procedure program (which can be used with the SPSS-X mainframe package) and a stand-alone program suitable for both mainframe and microcomputers.

The Mokken scale model is a stochastic elaboration of the well-known deterministic Guttman scale (Mokken, 1971; Mokken & Lewis, 1982; Mokken, Lewis, & Sytsma, 1986). The Mokken model is applied to dichotomous items for which one alternative is designated as positive with respect to the latent variable of interest (attitude, ability, etc.). This model belongs to the general family of the popular item response theory models (Mokken & Lewis, 1982). The Mokken model has been shown to be very useful for scaling social attitudes, political knowledge, political participation, and political efficacy (Gillespie, Ten Vergert, & Kingma, 1987a, 1987b; Lippert, Schneider, & Wakenhut, 1978; Mokken, 1969a, 1969b; Stokman, 1977; Stokman & Van Schuur, 1980). In psychological research, the Mokken model has been applied successfully to the construction of developmental scales of Piagetian concepts (Kingma, 1984; Kingma & Loth, 1985; Kingma & Reuvekamp, 1984; Kingma & Ten Vergert, 1985; Kingma & Van den Bos, in press), as well as to cumulative scaling of problem solving tasks (Henning, 1976) and sleep quality (Mokken & Lewis, 1982). The purpose of this paper is to describe a user procedure FORTRAN 77 program (which can be used with SPSS-X) and a stand-alone Fortran 77 program of

which both mainframe and MS-DOS versions are available. These programs, both named Mokscal, perform the Mokken scale analysis. Before presenting a review of the program and the various statistics, we describe in short-hand the underlying assumptions of the Mokken model, which are crucial for the appreciation of this rather unknown model.

## THE MOKKEN MODEL

The Mokken model treats the attitude, the ability, or any other variable of interest as a single latent trait on which the person's location is represented by the parameter  $\theta$  and the item's location is represented by the parameter  $\delta$ . Given a reasonably unidimensional set of items—that is, one dominated by the latent variable measured—the person's parameter ( $\theta$ ) can be estimated by the number of items to which a person responds positively, and the item parameter ( $\delta$ ) can be estimated by the proportion of people who respond positively to that item. The former is referred to as the person's scale score and the latter as the item difficulty. Items that fit the Mokken model must meet the so-called assumption of double monotony.

### The Assumption of Double Monotony

Assume that  $n$  persons answer  $k$  dichotomous items  $X_1, \dots, X_k$ , where  $X_i = 1$  or  $0$  if a person's answer to item  $i$  is positive or negative, respectively. The Mokken model specifies the relation between item  $X_i$  and the latent trait in terms of an ICC, denoted by  $P(X_i | \theta_j)$ . As formal ex-

---

Preparation of this paper was supported by Grant A1235 from the Natural Sciences and Engineering Research Council of Canada and by Grant H56-314 from the Netherlands Organization for the Advancement of Pure Scientific Research (Z.W.O.). Requests for reprints may be addressed to Terry Taerum, Department of Computing Services, 319C General Services Building, University of Alberta, Edmonton, Alberta T6G 2E1, Canada, or to Johannes Kingma, Department of Educational Psychology, University of Utah, Salt Lake City, Utah 84112.

pression indicates, this curve represents the probability of a positive response on item  $X_i$ , given respondent  $j$ 's location ( $\theta$ ) on the latent trait. An important feature of the Mokken model is that, unlike the parametric latent trait models (Hulin, Drasgow, & Parsons, 1983), it makes no assumption about the functional form of the item characteristic curve (ICC). Consequently, the Mokken model is referred to as a nonparametric model, and the resulting scale scores and item difficulties constitute ordinal rather than interval or ratio values.

Instead, the only constraint that the Mokken model puts on the ICCs is referred to as the *assumption of double monotony*. The first requirement of this assumption is that for any item in a Mokken scale, the probability of a positive response increases as  $\theta$  increases. To put this more formally, for any two persons  $i$  and  $j$  where  $\theta_i < \theta_j$ , the probability of a positive response on any item in the scale is less for person  $i$ . This requirement has been defined as *monotone homogeneity* (Mokken & Lewis, 1982; Molenaar & Sytsma, 1984), which is a necessary precondition for unidimensional measurement.

The second requirement to fulfill the assumption of double monotony is that for any value of  $\theta$ , the probability of a positive response decreases with the difficulty of the item. This means that the order of item difficulties remains invariant over the values of  $\theta$ , or, stated graphically, the ICCs do not intersect. Given these two requirements, it becomes possible to define unambiguously the difficulty of an item with a probability of .50. To determine whether a set of items forms a Mokken scale, two test procedures are performed.

### Testing the Monotone Homogeneity of a Set of Items

To test the requirement of monotone homogeneity, Mokken developed three related coefficients of scalability, which are based on the assumption of marginal independence (Mokken, 1971). The first,  $H_{ij}$ , measures the homogeneity or association between each pair of items. The second,  $H_i$ , measures the homogeneity of a particular item with respect to all other items and is obtained by aggregating across the coefficients of the relevant item pairs. The third,  $H$ , measures the homogeneity of the scale as a whole by aggregating across the coefficients of the individual items.

According to Loevinger (1948), the coefficient for the homogeneity of an item pair,  $H_{ij}$ , essentially measures the association in the  $2 \times 2$  table that is obtained by cross-classifying the two items (see Table 1). Let item  $i$  define the rows and item  $j$  define the columns, where item  $i$  is more difficult than item  $j$ . Under Guttman's deterministic model, we expect the top-right cell, the *error cell*, to be empty [i.e.,  $f(1,0) = 0$ ]. Under the model of statistical independence (or no association between  $i$  and  $j$ ), we anticipate that the expected value of the frequency of the error cell, defined as  $e(1,0)$ , equals the product of marginal frequencies divided by the sample size [i.e.,  $e(1,0) = f(1,.)f(.,0)/f(.,.)$ ]. Given in Equation 1,  $H_{ij}$ , the es-

**Table 1**  
**The Cross-Tabulation of Two Items**

Response to Item $i$	Response to Item $j$		Row Total
	1	0	
1	$f(1,1)$	$f(1,0)$	$f(1,.)$
0	$f(0,1)$	$f(0,0)$	$f(0,.)$
Column Total	$f(.,1)$	$f(.,0)$	$f(.,.)$

Note—Item  $i$  is assumed to be more difficult than item  $j$ . A 1 denotes a positive response; a 0 denotes a negative response.

timate of the index of item pair homogeneity, measures the proportional difference between cell frequency of the error cell expected under marginal independence and the actual cell frequency:

$$H_{ij} = \frac{[e(1,0) - f(1,0)]}{e(1,0)}, \tag{1}$$

where  $e(1,0) = f(1,.)f(.,0)/f(.,.)$  and  $i < j$ .

Readers familiar with the convention of using letters  $a, b, c$ , and  $d$  to represent the cell frequencies of a  $2 \times 2$  table may find this notation more convenient:

$$H_{ij} = (ad - bc)/(a+b)(b+d). \tag{2}$$

They may also notice the similarity between this index and other measures for  $2 \times 2$  tables. When items are independent,  $H_{ij}$  will be zero; when the error cell is empty,  $H_{ij}$  will equal unity.

The estimate of the coefficient of item homogeneity,  $H_i$ , is given in Equation 3:

$$H_i = \frac{\sum_{j=1}^k e_{ij} - \sum_{j=1}^k f_{ij}}{\sum_{j=1}^k e_{ij}}, \tag{3}$$

where  $i \neq j$ . The estimate aggregates the observed and expected frequencies for the error cell for each  $2 \times 2$  table that cross-classifies item  $i$  with the other items in the scale.  $H_i$  is analogous to the item-total correlation used in reliability analysis (Sytsma & Molenaar, 1987). It will be zero when item  $i$  is independent of the other items used in the scale. It will attain a maximum value of unity when the error cell in each of the relevant  $2 \times 2$  tables is empty. In terms of Equation 2, the numerator of Equation 3 consists of the sum of the differences between the diagonal and off-diagonal cross-product terms for the  $2 \times 2$  table. The denominator of Equation 3 sums the product of the appropriate marginals for these tables.

The estimate of the coefficient of scale homogeneity,  $H$ , is given in Equation 4, which aggregates the observed and expected error frequencies used to calculate  $H_{ij}$  for all item pairs:

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^k e_{ij} - \sum_{i=1}^{k-1} \sum_{j=i+1}^k f_{ij}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^k e_{ij}}, \tag{4}$$

where  $i = j-1$ . Again in terms of Equation 2, the numerator of Equation 4 consists of the sum of the differences between the diagonal and off-diagonal cross-product terms for all  $2 \times 2$  tables. The denominator of Equation 4 sums

the products of the appropriate marginal rows of these tables.  $H$  will be zero when all items are mutually independent, and it will attain the maximum value of unity when the error cells for all the  $2 \times 2$  tables are empty.

An asymptotic sampling theory of the estimates of the scale coefficients,  $H$  and  $H_i$ , has been developed (Mokken, 1971, pp. 157-160). On the basis of this theory, the null hypothesis of random response, also called the null case ( $H=0$  or  $H_i=0$ ), can be tested. A statistic called DELTA STAR has been derived from the estimate of the scale coefficient  $H$ , and on item level the estimate of  $H_i$  has been employed (see Mokken, 1971). The statistic DELTA STAR allows us to test the more specific null hypothesis of random response:

$$H_0 : E(H) = E(\text{DELTA STAR}) = 0.$$

The null hypothesis of random response is accepted when the numerical value of DELTA STAR is smaller than some critical value  $c$  defined by the researcher, where  $c$  is the  $z$  score from the standard normal distribution, which is obtained as follows:

$$c = z \text{ for } \alpha_1^* = \frac{\alpha}{\frac{1}{2}k(k-1)}, \quad (5)$$

where  $\alpha$  is the confidence level chosen by the researcher and  $\frac{1}{2}k(k-1)$  is the number of item pairs. When DELTA STAR is greater than  $c$ , the null case is rejected; that is, it may be concluded that the coefficient is not based on random response. Using this test on item level—that is, testing  $H_i$  against the  $c$  criterion—enables the researcher to weed out those items that are based on random response (or chance only).

The test of the null case is the first step toward deleting items that do not satisfy the Mokken scaling model. However, the researcher may also impose (a priori) additional requirements on the coefficients  $H_i$  and  $H$  in such a way that both  $H_i$  and  $H$  must exceed a numerical value  $u$  (where  $0 < u < 1$ , and  $u$  will be gradually increased, e.g., from .3 to .5). Only those items are selected of which the  $H_i$  exceeds this a priori criterion (and consequently the scale coefficient  $H$  of the selected items will also be greater than the critical value). The possibility of gradually imposing stronger requirements on the items is especially worthwhile for the construction of the stochastic Mokken scale in new research areas (Kingma & Ten Vergert, 1985).

The coefficients of scale and item homogeneity allow the researcher to judge whether the scale or the scalable set of items as a whole and the scalability of the individual items satisfy the established criteria. Mokken (1971) established a set of criteria for the coefficients  $H$ ,  $H_i$ , and  $H_{ij}$  in order to analyze the homogeneity of a scale. First, all  $H_{ij}$  should be greater than zero. Second, all  $H_i$  (and thus all  $H$ ) should be greater than a predetermined constant  $u$ . The following classification of scale has frequently been used (Kingma, 1984):  $.50 \leq H$ , a strong scale;  $.40 \leq H < .50$ , a scale of medium strength; and  $.30 \leq H < .40$ , a weak scale.

In practice, when an item does not meet these criteria, it is eliminated from the scale. Subsequently, the coefficients  $H_{ij}$ ,  $H_i$ , and  $H$  are recomputed for the remaining subset of items, and a new inspection has to be performed whether or not these coefficients meet the criteria. This process is repeated until a sufficiently strong Mokken scale is obtained.

### Monotonicity in the Item Difficulties

The estimates of the coefficients of scalability ( $H$ ,  $H_i$ , and  $H_{ij}$ ) are used to test the monotone homogeneity of the items. This test is a necessary, but not sufficient, condition to meet the assumption of double monotony underlying the Mokken model. The test of the monotonicity of item difficulties, together with the test for monotone homogeneity of the items, provides a sufficient condition to test the assumption of double monotony.

The test of the monotonicity of the item difficulties involves an inspection of the  $P$  and  $P_0$  matrices, which contain the probabilities of two positive and two negative responses, respectively, to all possible pairs of items. This test is based on another assumption of the Mokken model, namely the assumption of local stochastic independence of item responses for a fixed subject parameter value. More specifically, item responses are conditionally independent, given the same value of  $\theta$ , so that the conditional probability of joint responses for persons with the same value of  $\theta$  equals the product of the marginal probabilities of these responses. The test of monotonicity in item difficulties analyzes the observable unconditional probabilities of each item pair in the  $P$  and  $P_0$  matrices. Since local independence is assumed for a given  $\theta$  value, the unobservable conditional probabilities imply the observable unconditional probabilities (see Mokken, 1971).

For example, when items 1, 2, and 3 represent decreasing difficulty, the (conditional) probability of a pair of positive responses will be the greatest for items 2 and 3, next greatest for items 1 and 3, and finally for items 1 and 2. Similarly, the (conditional) probability of a pair of negative responses will be the greatest for 1 and 2, followed by 1 and 3, and 2 and 3.

In the  $P$  and  $P_0$  matrices, the rows and columns are ordered, respectively, from left to right and from top to bottom according to decreasing levels of item difficulty. To meet the criterion of monotonicity in item difficulties, the probability of a pair of positive responses should increase in the  $P$  matrix from left to right in each row and from top to bottom in each column. Similarly, the probability of a pair of negative responses should decrease in the  $P_0$  matrix. The statistical significance of departures from this pattern can be tested by means of a simple runs test (Molenaar, 1983a). In practice, those items that violate significantly the criterion of monotonicity in item difficulties are removed from the scale, and the analysis is repeated until no significant disturbances are observed in both matrices.

A Mokken scale has been found when the items meet the criterion of double monotony; that is, the data must be in agreement with the assumption of monotone homo-

generity, as well as with the assumption of monotonicity in item difficulties.

**A Test for Robustness of the Scale as Goodness of Fit Criterion**

Once the researcher finds a stochastic Mokken scale, he/she can test the robustness of the scale (the non-null case) across different subgroups of the sample. The sample may be divided into groups according to different external criteria (e.g., age, sex, religion, and socioeconomic status). If such an external criterion (e.g., socioeconomic status) is adopted, one may test the null hypothesis that the coefficients of scalability have the same values for each subgroup. Therefore, the probability distribution of  $H$  and  $H_i$  must be estimated for these subsamples by means of Equations 3 and 4. Mokken (1971, pp. 164–167) showed that for large  $n$ , the quantity  $(H_n - H)/S(H_n)$  is approximately standard normally distributed, where  $S^2(H_n) = 1/n H_n^2 \sum_n H_n$  is the estimated asymptotic variance of the subsample, and  $H_n$  is the estimate of the scale coefficient for subgroup  $n$ , and  $S(H_n)$  is the estimated asymptotic standard deviation of the subsample scale coefficient. The mean value  $\bar{H}$  for the different sample coefficients  $H_h$  is defined as follows:

$$\bar{H}_h = \frac{\sum_{h=1}^p H_h / S^2(H_h)}{\sum_{h=1}^p 1 / S^2(H_h)}, \tag{6}$$

The statistic  $T$  is used to test the null hypothesis that the values  $H_h$  for subgroups  $h$  ( $h=1, 2, \dots, p$ ) are equal ( $H_1 = H_2 = \dots = H_p$ ).  $T$  has been defined as follows:

$$T = \sum_{h=1}^p \frac{(H_h - \bar{H})^2}{S^2(H_h)}, \tag{7}$$

Under the null hypothesis,  $T$  has approximately a chi-square distribution with  $p-1$  degrees of freedom. If the null hypothesis cannot be rejected (i.e., if  $T$  is smaller than a critical value dependent on the number of the degrees of freedom and on a significance level  $\alpha$ ), the mean value for  $\bar{H}_h$  (i.e.,  $\bar{H}$ ) can be considered a pooled estimate  $H_h$  (now supposed all equal) for the  $p$  samples. The samples or subgroups must be disjoint.

The Mokken test for robustness is considered a goodness of fit test, for an internal criterion (e.g., score groups; see Molenaar, 1983b) as well as for an external criterion (i.e., when the scale coefficient remains equivalent for the different subgroups). This test is comparable to the Anderson goodness of fit test in the one-parameter item response model, or the Rasch model (Anderson, Kearny, & Everett, 1968; Gustafson, 1980; Molenaar, 1983b).

In summary, the Mokken test provides a statistical basis to decide whether the ordering of a set of items is invariant across different subpopulations. If the assumption of double monotony holds for a population, the order of the item difficulties should remain invariant (subject to sampling error) for any sample from the population, whereas the order of the respondents (based on scale scores) will remain invariant to different samples of items

(again, subject to sampling error). This feature is referred to as *specific objectivity* (Rasch, 1960).

The Mokken test can also be performed on item level by replacing  $H_{ih}$  for  $H_h$  and  $S(H_{ih})$  for  $S(H_h)$  in Equations 6 and 7.

**Confidence Intervals**

Since the Mokken robustness analysis provides estimates for both  $S(H)$  and  $S(H_i)$  for the sample data, a confidence interval can be asymptotically determined for  $H$  and each  $H_i$ . Let  $d$  be the  $z$  score from the standard normal distribution with two-sided probability  $\alpha$ . For a large sample size  $n$ , an approximate confidence interval for the population coefficient  $H$  at probability level  $1/2\alpha$  is

$$H - dS(H) \leq H \leq H + dS(H), \tag{8}$$

where  $H$  is obtained with Equation 4. Similarly, the confidence interval of the population coefficient  $H_i$  for item  $i$  can be obtained by substituting  $H_i$  for  $H$  and  $S(H_i)$  for  $S(H)$  in Equation 8.

**Reliability of Test Scores in Mokken's Nonparametric Item Response Model**

Mokken's approach to reliability estimation avoids the assumptions of item equivalence that are commonly employed in classical test theory. The assumption of double monotony permits an estimation method for the reliability coefficient  $\rho_{xx}$ , which is not based on replication or internal consistency. In the  $P$  matrix, the nondiagonal cell entries contain the probabilities of the positive responses to each item pair. If we could estimate the values of the diagonal elements of this  $P$  matrix, we would have an approximation for all items of their probability of a positive response on two different test occasions. This probability is denoted by  $f_{(i,i)}$ . Mokken devised two methods to estimate these diagonal elements from off-diagonal elements of the  $P$  matrix. The estimate of  $f_{(i,i)}$  is defined for the method of extrapolation as

$$\tilde{f}_{(i,i)} = \frac{f_{(i,i)} f_{(i,i+1)}}{f_{(i,i+1)}}, \tag{9}$$

where  $f_{(i,i)}$  and  $f_{(i,i+1)}$  are, respectively, the sums of the probabilities of the  $i$ th and the  $i+1$  columns of the  $P$  matrix, and  $f_{(i,i+1)}$  is the  $(i,i+1)$ th off-diagonal element.

When  $f_{(i,i-1)}$  is taken as the off-diagonal element, then the method of extrapolation defines the estimate of  $f_{(i,i)}$  as follows:

$$\tilde{f}_{(i,i)} = \frac{f_{(i,i)} f_{(i,i-1)}}{f_{(i,i-1)}}. \tag{10}$$

According to the method of interpolation, the estimate of the diagonal element  $f_{(i,i)}$  is defined as

$$\tilde{f}_{(i,i)} = f_{(i,i-1)} + \frac{(f_{(i,i)} - f_{(i,i+1)})(f_{(i,i+1)} - f_{(i,i-1)})}{(f_{(i,i+1)} - f_{(i,i-1)})}. \tag{11}$$

The method of interpolation is used for  $i = 2, 3, \dots$ ,

$k-1$ , whereas the method of extrapolation has to be employed for  $i=1$  and  $i=k$ . Having found the estimates  $f_{(i,i)}$  according to one of these two methods, the approximate reliability coefficient  $\rho$  can be computed, where  $\rho$  is defined as

$$\tilde{\rho} = \left[ \sum_{i=1}^k (\tilde{f}_{(i,i)} - f_{(i,i)}^2) + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^k (f_{(i,i+1)} - f_{(i,i)} f_{(i,i+1)}) \right] / \left[ \sum_{i=1}^k f_{(i,i)} (1 - f_{(i,i)}) + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^k (f_{(i,i+1)} - f_{(i,i)} f_{(i,i+1)}) \right], \quad (12)$$

and where  $0 \leq \rho \leq 1$ . Note that the reliability coefficient is calculated only for a set of items and not for the individual items. Some further improvements of estimating the reliability coefficient from the estimates of diagonal elements have been made (Molenaar & Sytsma, 1984; Sytsma & Molenaar, 1987). By interchanging the positive responses for the negative responses, the direction of the Mokken scale is reversed. The estimates of diagonal elements of the  $P$  matrix are defined as follows. For the two procedures of the method of extrapolation, the estimates of the diagonal elements of the  $P$  matrix are, respectively,

$$\tilde{f}_{(i,i)} = \frac{f_{(i-1,i)}(1-f_{(i,i)})}{(1-f_{(i,i-1)})} + \frac{f_{(i,i)}(f_{(i,i)}-f_{(i,i-1)})}{(1-f_{(i,i-1)})} \quad (13)$$

and

$$\tilde{f}_{(i,i)} = \frac{f_{(i,i+1)}(1-f_{(i,i)})}{(1-f_{(i,i+1)})} - \frac{f_{(i,i)}(f_{(i,i+1)}-f_{(i,i)})}{(1-f_{(i,i+1)})}. \quad (14)$$

The method of Molenaar and Sytsma (1984), called the MS method of estimating reliability, consists of substituting the average of the estimates of the diagonal elements from Equations 9, 10, 13, and 14 into Equation 11.

## PROGRAM DESCRIPTION

The Mokken scale user procedure program (for SPSS-X) and the stand-alone program (for both mainframe Ahmdahl and IBM PC/AT microcomputers using the MS-DOS operating system) were written in FORTRAN 77, are double precision, and occupy 268K of memory. With minor modifications, these programs can be operational on other microcomputers or mainframes.

### Input

The user procedure program for SPSS-X and the stand-alone program, both called Mokscal, have the same command file structure. The following is an example of this command file:

Record No.

- 1 Title 'Attitude data 1987'
- 2 File handle atti/name = 'mokken1.dat'
- 3 Data list file = atti
- 4 / item1 to item20 1-20 group 21

- \*5a Value labels item1 1 'Test1' 2 'Test2'
- \*5b / group 1 'Male' 2 'Female'
- 6 Userproc Name = Mokken
- 7 / item2 item4 to item7 item9 to item20
- \*8 / Select item2 item4 item18 to item20
- 9 / Options positive = 2 alpha = .01  
lowerbound = 0.4 type = search (or test)  
(robust)
- 10 print = P, P<sub>0</sub>, H, DELTA

On the first record of the command file, the researcher describes the title he/she wants to give the job. The second record specifies the file handle. Note that 'mokken1.dat' refers to an ASCII file containing the data, whereas the file name atti is a local file name necessary for the data list. The third record concerns the data list in which the local file specification, file = atti, is identical to that denoted in the file handle of record 2. Record 4 is a continuation of the data list; it contains the specifications as to which items (variables) the program Mokscal has to read from the data file mokken1.dat. The program knows where to read the data; each item 1 to 20 occupies one column, starting in column 1. The variable group is located in column 21 of each record in the data file. In this example, it is clear that there is only one record per subject in the data file mokken1.dat. When a data file contains more than one record per subject and the items for the Mokken scale analysis are scattered on different records, the number of slashes indicates how many records the program Mokscal has to read per subject. Suppose a data file contains five records per subject, the data list specification may be:

- ```
/ var1 to var4 3-6
/ var5 to var6 18-19
/ var7 to var8 1-2 var9 to var12 21-24
/ var13 1
/ var14 4 var15 12
```

On the first record, each variable 1 to 4 occupies one column: variable 1 in column 3, variable 2 in column 4, and so forth. On the second record, variables 5 and 6 are located, respectively, in columns 18 and 19. The third, fourth, and fifth slashes (always in column 2 of the command file) refer, respectively, to the third, fourth, and fifth records. The location of the variables on these records is self-evident based on the previous description.

Suppose that the researcher wants to perform a Mokken scale analysis on a data file that contains six records per subject, but the variables are located on records 4 and 6. Then the data list specification may be:

- ```
///
/ var1 to var6 1-6 var7 to var10 21-30
// var11 to var20 1-10
```

The program Mokscal skips the first three records for each subject, and then it reads the variables 1-6 followed by variables 7-10 from the specified columns in record 4. Subsequently, record 5 will be skipped, and variables

11-20 are read from record 6 according to their column specifications.

On many occasions, the researcher may use a spreadsheet program for data entry in which the variables are separated by blanks, commas, or so forth. The program Mokscal is able to read this type of data, because it also accepts FORTRAN format statements in the data list specification. Assume that a set of data has been entered by means of a spreadsheet program, there is one record per subject, and 20 variables have to be read. The variables are separated by blanks and variable 1 is located in column 1. Then the data list specification may be:

```
/ v1 to v20 (f1.0, 19f2.0).
```

When the spreadsheet program delivers commas or other symbols to separate the variables, the format statement for this example for the data list specification is:

```
/ v1 to v20 (f1.0, 19(x(f1.0)))
```

In summary, the data list specification enables the researcher flexible use of the program Mokscal, without the need to alter the original (ASCII) data file for different Mokken scale analyses.

The fifth record in our example command file is optional (thus the asterisk). It specifies the value label for a particular variable (i.e., a name is attached to a particular variable). These value label names (descriptive labels) are printed on the output instead of the variable name, as specified in the data list specification on record 4. The use of value labels often makes the output more readable. In our example command file, the value label attached to item 1 when item 1 = 1 is Test1; when item 1 = 2, the value label is Test2. The value label for each variable is separate from the value label for the preceding variable by a slash. For example, a slash precedes the value label assignment for group on record 5b.

Record 6 of the command file specifies that after the data are read (either by SPSS-X or by the stand-alone program read shell), the user procedure part of the Mokscal program starts.

On record 7, the researcher may specify which variables are included in the actual Mokken scale analysis. In our example, command items 2, 4-7, and 9-20 are included for the scale analysis.

Record 8 is used only when the extension of an existing Mokken scale is the focus of interest. Note that this slash select statement is employed only with the option = type = search statement (see record 9). In other cases, the select record is not included in the command file. Suppose that a researcher has found in previous research that 5 items of interest form a nonparametric Mokken scale. Suppose also that the researcher wants to extend this scale by adding 15 new items to the 5 old ones. By means of the select statement of record 8, the researcher can specify the old items as a starting point of the scale analysis (in our example, command file items 2, 3, and 18-20). The program Mokscal searches through the set of 15 new items to determine whether the whole set or a subset of the new

items may form together with the 5 old items into a non-parametric Mokken scale.

Record 9 of the command file contains different options. Because the program Mokscal is for dichotomous items only, "options positive = 2" indicates that a score of 2 on every item is considered the positive alternative (the default value is positive = 1; i.e., when an option is not stated on the option list, the program Mokscal considers that a score of 1 is considered the positive alternative). The  $\alpha$  level for testing whether an item fits the scale has to be set to a constant  $c$  by the researcher. In our example command file,  $\alpha = .01$  (the default value is  $\alpha = .05$ ). The option "lowerbound" specifies the lowerbound of scale coefficient  $H$ , item coefficient  $H_i$ , and coefficient  $H_{ij}$  for each item pair  $i$  and  $j$ . The default value for the lowerbound is .50. The procedure to be used to analyze whether a set or subset of items form a Mokken scale is declared by the option "type." Type = search is used when the research is exploring whether a set or a subset of items may be considered a Mokken scale. Type = test is used for testing whether a set of items as a whole (as specified in record 7) will form a Mokken scale. When type = robust, the Mokken test for robustness is performed for the variables as specified in record 7. The group variable also has to be declared on this record. The researcher wants to test the robustness of the scale for both sex groups, record 7 of our example command file becomes:

```
/ item2 item4 to item7 item18 to item20 by group(2).
```

In record 10, print = narrow is employed when one wants to view the output on the console (i.e., the output will have a record length of 80 characters). This option may also be used when printers of microcomputers do not have a condensed mode for printing a record length of 132 characters. The width default value has been set to 132 characters. The option "print" enables the researcher to specify which part of the output will be printed or written to a file; however, in exploring a set of items, one may bypass the print of all output. For example, when one is interested only in the  $H$  coefficients, the print option may be set to print=H.

The command file is employed in the common batch procedures for SPSS-X jobs when the user procedure file of Mokscal is used. The stand-alone program is started by calling Mokken (\$run Mokken on MTS operating system; Mokken, followed by a carriage return, on microcomputers). Subsequently, the program asks for the name of the command file. Following this, the program asks for the name of the output file or output device.

The program employs different types of algorithms for the search test, and robustness test procedures.

### Search Procedure

When exploring a set of items, the researcher may be interested primarily in constructing a new stochastic Mokken scale that (on the basis of the content) is thought to be more or less homogeneous. The logic behind such a

search procedure runs as follows: a selection of items has to be made (the number of items as large as possible) that meet the criteria of scalability as described previously. In addition to the statistical criteria, the approach of the search procedure is a familiar one in statistical methods aiming for optimal selection of  $p$  items from a set of  $k$  items. In theory, such selection is a very simple procedure, because for given  $p$ , all possible  $p$ -tuples from the set of  $k$  items have to be inspected. However, such a selection procedure is rather inefficient (tedious and time-consuming), because astronomically great numbers of  $p$ -tuples may be involved.

The program Mokscal provides an alternative selection procedure. The method developed is a straightforward maximization of the scale coefficient  $H$ . The maximization procedure consists of a stepwise technique of constructing a scale from a set of  $k$  items. The lowerbound of  $H$  is set to the critical value  $c$  (see input description). The program computes the cross-tabulations ( $2 \times 2$  tables) for all item pairs (see Table 1). The maximization procedure starts with the choice of the best pair of items. This choice is based on the following criteria: (1) the coefficient  $H_{ij}$  must be significantly greater than zero at a predefined level of confidence  $\alpha$  (see input description); (2)  $H_{ij}$  is the greatest for all possible item pairs that meet criterion 1; and (3)  $H_{ij}$  must be equal to or greater than the chosen lowerbound  $c$  (see input description).

The algorithm used in the program Mokscal ensures uniqueness of the selection of the item pair when more item pairs are equally eligible. In this circumstance, the starting pair of items contains the most difficult items in the sequence of sample difficulties. If no unique pair can be found, the pair of next most difficult items is chosen. Following the choice of an optimal starting pair of items, single items are added to the scale in consecutive steps. In general, this recursive process of adding an item to the scale can be described as follows: Let two items be selected from a set of  $r$  items as a starting pair for the new scale. Then from the remaining set,  $(r-2)$  items the  $r$ th item ( $i_r$ ) will be added to that scale only when it meets the following three criteria: (1) Item  $i_r$  should have a positive correlation with each item that has already been selected for the scale. An item that correlates negatively with one of these already scalable items is rejected by the program in the consecutive steps for the construction of the new scale at hand. (2) The item coefficient of scalability,  $H_{i_r}$ , of item  $i_r$ , with respect to the already selected items should be significantly greater than zero at the selected  $\alpha$  level (see input description), according to the one-sided test based on the statistic DELTA STAR. (3) The coefficient  $H_{i_r}$  with respect to the already selected items should be greater than, or equal to, the critical lowerbound  $c$  (see input description).

The three preceding criteria are employed by the program Mokscal to determine the set or subset of items that are in agreement with the definition of a Mokken scale. In addition, the following three criteria are employed by the program to assure that the best item is chosen from

the set of items, as well as that the choice is unique: (4) The scale consisting of the already selected items and item  $i_r$ , should have the highest value  $H$  relative to any other item from the set of remaining items. (5) When a unique choice is impossible, because some items may meet criteria 1 to 3 and they may also have equal maximum  $H$  values, then the item  $i_r$ , that has the greatest  $H_{i_r}$  value with respect to the already selected items is chosen by the program. (6) When more than one item still can be chosen, then the most difficult is chosen by the program Mokscal (i.e., the item with the lowest  $p$  value, or proportion correct responses) in the sequence of sample difficulties.

At the end of this search procedure, the resulting scale has to be evaluated on the basis of the  $P$  and  $P_0$  matrices. In the  $P$  matrix, the proportions of correct responses to each item pair  $i$  and  $j$  should increase from left to right in each row and from top to bottom in each column. In the  $P_0$  matrix, the proportions of incorrect responses to each item pair  $i$  and  $j$  should decrease from left to right in each row and from top to bottom in each column. On the basis of a simple run test (see Molenaar, 1983a), the program Mokscal marks those cell entries in which the proportion differs significantly from the expected pattern. When significant disturbances are observed in either the  $P$  or the  $P_0$  matrices, the set of items selected by the search procedure may not be considered a stochastic Mokken scale, because they do not meet one of the assumptions of double monotony, that is, the monotony in item difficulties. In practice, the researcher deletes items that are responsible for these disturbances (by altering the specification of the items to be searched in record 7 of the command file), and he/she repeats the search procedure for these remaining items. When no significant disturbances are observed in either the  $P$  or the  $P_0$  matrices, it may be concluded that these constitute a stochastic Mokken scale, because they meet the criterion of double monotony (the assumption of monotone item homogeneity has already been met by those items selected during the search procedure, and the lack of disturbances in either the  $P$  or the  $P_0$  matrices implies that the items of the scale also meet the assumption of monotonicity in item difficulties).

In addition, the program Mokscal does not stop when a final scale is established for a subset of items. The program also analyses whether the remaining items (that do not fit the first final scale) constitute a separate scale.

The search procedure is ended when no items in the remaining set meet criteria 1 to 3, or when no more items are available, or when the set of items eligible in terms of criterion 1 is empty.

#### Extension of an Existing Mokken Scale

Once a Mokken scale has been found and its usefulness has been shown, it might be worthwhile to extend the scale. Improvements and extensions of these scales may be made by adding new items and trying to vary difficulty levels for better differentiation. For this purpose, the search procedure can be modified. Instead of

beginning with the selection of the initial best pair, items are selected as the start set, and, subsequently, the search procedure of adding new items starts as previously described. The Mokken scale program performs this extension procedure when the following record is added in the command file before the options (see input description):

```
/SELECT var2 var5 var6 ... vark
```

where var2, var5, var6, and vark are the items of the original scale (in the options type = search). Those new items that meet the first three criteria of the search procedure are added to the original scale. Further inspection of the *P* and *P*<sub>0</sub> matrices must show that these new selected items meet the assumptions of double monotony. Items that violate these requirements have to be deleted, and the procedure of the search for the extension has to be repeated for the remaining set of items. As in the SEARCH procedure, the Mokken scale program searches the items that do not fit on the extended scale to determine whether these items (or subsets) may form another scale.

**Level of Significance in the Search Procedure**

In the search procedure of the Mokken scale program, a number of statistical tests are performed at each step to determine the ultimate selection of an item. It may be argued that due to repeated testing, the alpha risk may increase. Consequently, the obtained results may have a very low probability of being repeated in other experiments.

To avoid this risk of capitalizing on chance, the sample may be randomly split into two halves to replicate the analysis; however, the power of the tests will be reduced. The Mokken scale program offers an alternative; that is, a gradually increasing protection level is employed to moderate the effects of possible chance results by imposing a continual reduction of the level of significance in the consecutive steps. This algorithm runs as follows: First at the beginning of the search procedure, a level of significance (mostly  $\alpha = .05$  or  $.01$ ; see input description) is chosen. Second, from an initial set of *k* items, the test for the selection of the best starting pair is performed at the actual level of significance given in Equation 5:

$$\alpha_1^* = \frac{\alpha}{\frac{1}{2}k(k-1)},$$

where the denominator indicates the number of comparisons involved. Third, after eliminating those items that do not meet the first criterion of the search procedure (i.e., positive correlation), it is supposed that the remaining set of items consists of *f*<sub>2</sub> items. The tests necessary for the selection of the third item are performed at level of significance:

$$\alpha_2^* = \frac{\alpha}{\frac{1}{2}k(k-1) + f_2}. \tag{15}$$

Fourth, the significance (protection) level used in the Mokken scale program can be generally defined for *g*

items being selected for the scale (from the set of *k* items), with *f*<sub>2</sub> items being retained in the remaining pool (after eliminating the negatively correlated items). The tests necessary for the selection of the (*g* + 1)th item are then computed by the program at significance level:

$$\alpha_{(g+1)}^* = \frac{\alpha}{\frac{1}{2}k(k-1) + \sum_{j=2}^g f_j}. \tag{16}$$

**The Evaluation of a Whole Set of Items as One Scale**

Suppose it has been shown in previous research that a set of *k* items forms a stochastic Mokken scale. When this set of items is used in another study, the Mokken scale program may be employed to evaluate whether the whole set of *k* items again forms a Mokken scale (under the same restrictions as in the earlier study, i.e., the same alpha level and lowerbound; see input description) by setting type = test (record 9 of the command file; see input description). The Mokken scale program evaluates whether the whole set of items forms one scale (i.e., no search procedure is performed). The algorithm at the basis of this test procedure computes the intercorrelations of all item pairs, the item coefficients *H*<sub>*i*</sub>, and the scale coefficient *H*. Two statistical levels of evaluation can be distinguished in this test procedure. First, the by-the-program computed values of the coefficients *H*<sub>*i*</sub> and *H* are tested one-sided against the null hypothesis of random response using the statistic DELTA STAR. DELTA STAR must exceed some critical *z* value at an  $\alpha$  level of

$$\alpha_{crit} = \frac{\alpha}{\frac{1}{2}r(1-r)}, \tag{17}$$

where  $\alpha$  in the numerator is the confidence level defined by the researcher (see input description) and *r* is the number of items. The critical *z* value is derived by the program from a table of the standard normal distribution for the computed  $\alpha_{crit}$ . In addition to this procedure, it is evaluated that no item is negatively correlated. The second step involves the evaluation of whether these coefficients exceed the lowerbound value *c* (see input description). When the whole set of items satisfies these two steps, the hypothesis of scalability is accepted. However, when one or more of the *H*<sub>*i*</sub>'s coefficients have negative values or values that do not differ significantly from zero, the hypothesis of the scalability of the set of items as a whole is rejected.

**Test for Robustness of the Mokken Scale**

The robustness (or invariance) of a found Mokken scale for different samples (or subsamples) can be tested by means of the Mokken test. The group variable has to be defined in record 7 and type = robust in record 9 (see input description).

The Mokken test option computes the *T* statistic for both the scale as a whole and for each of the items. *T* has an approximate chi-square distribution with *p* - 1 degrees of

freedom (where  $p$  is the number of groups). When  $T$  exceeds the critical value, the scales may not be considered invariant across the groups. The  $T$  statistics on item level provides the researcher information about which items are responsible for such a significant difference. By deleting those items with significant  $T_i$  values (in a stepwise manner) and subsequently rerunning the program with the option type = robust, a scale may be obtained that is robust for the different subgroups (for a practical application of this approach, see Gillespie et al., 1987a; Kingma & Ten Vergert, 1985).

### Output

The Mokken scale program (both the user procedure file for SPSS-X and the stand-alone programs for main-frame and microcomputers) provides the following output for the search procedure (type = search): (1) A description of the command file is given, followed by a specification of which columns and records are read in the data file. (2) An example of the syntax used in the command file is printed, enabling the researcher to change very quickly an error made. (3) A specification of the items used in the actual analysis are displayed and the number of cases read from the data file; (4) The  $P$ ,  $P_o$ , and  $H_{ij}$  matrices are given for all items; (5) Four reliability coefficients are provided for the whole set of items: one is based on Mokken's method of extrapolation to find the diagonal elements of the  $P$  matrix, the second is based on Mokken's method of interpolation, the third is the  $MS$  reliability coefficient (Molenaar & Sytsma, 1984; Sytsma & Molenaar, 1987, and the last is the Kuder Richardson coefficient of homogeneity (the  $KR_{20}$  and the item-corrected total correlations). (6) The phi inverse coefficient—that is, the  $z$  value—for the best start pair, the scale coefficient  $H$ , and the scale DELTA STAR statistic, as well as the actual  $\alpha$  level used, are given. On item level  $H_i$  coefficients, the statistics DELTA STAR and the  $p$  values (proportion correct responses) are given.

In the second step of adding an item to the scale, the output consists of (1) the scale coefficient  $H$  for the three items, the statistic DELTA STAR for the scale as a whole, and the actual  $\alpha$  level used; (2) on item level, the  $p$  values,  $H_i$  values for each of the three items, and their statistics DELTA STAR; and (3) the actual  $\alpha$  level used for testing and the phi inverse coefficient (i.e., the  $z$  score of a significance level; see Equations 14 and 15). For the third and following step, this intermediate output is given, so that for each step the changes in the item coefficients  $H_i$  and the statistics DELTA STAR can be evaluated.

When the search procedure of the Mokken scale program has found a final scale for the selection of items from the original set, the output is as follows: (1) the scale coefficient  $H$ , the statistic DELTA STAR of the found scale, the  $p$  values of the scalable items, their coefficients  $H_i$  and their statistics DELTA STAR, the phi inverse coefficient, and the actual  $\alpha$  level used; (2) the  $p$  values,

$H_i$  coefficients, and the statistics DELTA STAR of the non-scalable items that did not fit the final scale; (3)  $P$ ,  $P_o$ , and  $H_{ij}$  matrices and the four reliability coefficients for the final scale.

The search procedure is repeated by the program for those items that did not fit into the first final scale. The same output as described above is obtained for these items.

The procedure to extend an already existing Mokken scale (select = vari to varn and type = search) provides output that is almost identical to that of the search procedure. The only difference is that the output does not contain information about the start pair, but starts at that place in the output that contains the information about all items of the already existing scale, followed by the stepwise output for each new item added to this scale, as in the search procedures.

The output for the evaluation of a set of items as one scale (type = test) is as follows: (1) the  $P$ ,  $P_o$ , and  $H_{ij}$  matrices for all items; (2) the four reliability coefficients for the whole set of items; (3) the scale coefficient  $H$  and scale DELTA STAR; and (4) the  $p$  values, the  $H_i$  coefficients, and the statistics DELTA STAR for each individual item. If all  $H_i$  and  $H_{ij}$  meet the criteria (see search procedure), then the program stops. If some items are rejected, because they do not meet these criteria, the program provides additional output (identical to the search procedure) for both the scalable and unscalable items.

The following output is given for the test of the robustness of the found Mokken scale across different samples or subgroups (type = robust): (1) for all samples together, the description of the command file, the input format of the items specified, a list of the items used in the analysis, and the number of cases read from the data file; (2) a survey of the total scale scores for total population (all samples), the percentage and the different response patterns that constitute a particular total score (items are ordered in these patterns according to decreasing difficulty), and a survey for each sample or subgroup; (3) the  $P$ ,  $P_o$ , and  $H_{ij}$  matrices for all samples together and the four reliability coefficients; (4) the  $E$ ,  $E_o$ ,  $H$ ,  $S^2(H)$ , and the statistic DELTA STAR for both the scale and for each individual item level and the item difficulties; (5) the same output for each subsample; (6) the mean of the coefficient  $\bar{H}$ , the statistic  $T$ , the degrees of freedom, and the probability level for the whole scale, as well as for the individual items.

### Further Elaborations and Discussions on the Mokken Scale

The Mokken scale analysis is relatively unknown in North America. See Mokken (1971) for a thorough review of this nonparametric item response theory model. More recently, several discussions about the usefulness of the Mokken scale analysis have appeared in the literature (Jansen, Roskam, & Van den Wollenberg, 1984; Roskam, Van den Wollenberg, & Jansen, 1986; Sytsma, 1984,

1986). Further refinements of the model were reported by Molenaar (1982, 1983a) and Sytsma and Molenaar (1987). These refinements are incorporated in the present program.

### Limitations

The Mokken scale program is limited at present to the analysis of a maximum of 100 items. This limitation may be modified by altering dimensions of the arrays. The program accepts scores from an unlimited number of subjects.

## PROGRAM AVAILABILITY

Program listings of both the SPSS-X user procedure file and the stand-alone programs (both mainframe and micro-computer versions) can be obtained from either author on a 5.25-in. flexible disk (MS-DOS operating system) for a nominal fee. For the microcomputer, a compiled program is available for IBM PC/AT microcomputer or compatibles with math coprocessor. A user's guide and example input and output will accompany the listings of the source codes.

## REFERENCES

- ANDERSON, J., KEARNEY, G. E., & EVERETT, A. V. (1968). An evaluation of Rasch's structural model for test items. *British Journal of Mathematical Psychology*, *21*, 231-238.
- GILLESPIE, M., TEN VERGERT, E. M., & KINGMA, J. (1987a). Using Mokken methods to develop robust cross-national scales: American and West German attitudes toward abortion. *Social Indicators Research*, *19*, 74-95.
- GILLESPIE, M., TEN VERGERT, E. M., & KINGMA, J. (1987b). Using Mokken scale analysis to develop unidimensional scales: Do the six abortion items in the NORC GSS form one or two scales? *Quality & Quantity: International Journal of Methodology*, *21*, 1-15.
- GUSTAFSON, J. (1980). Testing and obtaining fit of data to the Rasch model. *Bulletin of British Psychological Society*, *33*, 205-233.
- HENNING, H. J. (1976). Die Technik der Mokken Skalenganalysen. *Psychologische Beiträge*, *18*, 410-430.
- HULIN, C. L., DRASGOW, F., & PARSONS, C. K. (1983). *Item response theory*. Homewood, IL: Dow Jones-Irwin.
- JANSEN, P. G. W., ROSKAM, E. E., & VAN DEN WOLLENBERG, A. L. (1984). Discussion on the usefulness of the Mokken procedure for nonparametric scaling. *Psychologische Beiträge*, *26*, 722-735.
- KINGMA, J. (1984). A comparison of four methods of scaling for the acquisition of early number concept. *Journal of General Psychology*, *110*, 23-45.
- KINGMA, J., & LOTH, F. L. (1985). The validation of a developmental scale for seriation. *Educational & Psychological Measurement*, *45*, 321-328.
- KINGMA, J., & REUVEKAMP, J. (1984). The construction of a developmental scale for seriation. *Educational & Psychological Measurement*, *44*, 1-23.
- KINGMA, J., & TEN VERGERT, E. M. (1985). A nonparametric scale analysis of the development of conversation. *Applied Psychological Measurement*, *9*, 375-387.
- KINGMA, J., & VAN DEN BOS, K. (in press). On the sequentiality in concept development. *Cognitive Systems*.
- LIPPERT, E., SCHNEIDER, P., & WAKENHUT, R. (1978). Die Verwendung der Skalierungsverfahren von Mokken und Rasch zur Überprüfung und Revision von Einstellungsskalen. *Diagnostica*, *24*, 252-274.
- LOEVINGER, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507-530.
- MOKKEN, R. J. (1969a). Dutch-American comparisons of the "sense of political efficacy." *Quality & Quantity*, *3*, 125-152.
- MOKKEN, R. J. (1969b). Dutch-American comparisons of the "sense of political efficacy": Some remarks on cross-cultural robustness of scales. *Acta Politica*, *4*, 425-440.
- MOKKEN, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton.
- MOKKEN, R. J., & LEWIS, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychology Measurement*, *6*, 417-430.
- MOKKEN, R. J., LEWIS, C., & SYTSMA, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement*, *10*, 279-285.
- MOLENAAR, I. W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden*, *8*, 145-164.
- MOLENAAR, I. W. (1983a). Rasch, Mokken en school beleving [Rasch, Mokken and school experience]. In S. Lindenberg & F. N. Stokman (Eds.), *Modellen in de Sociologie [Models in Sociology]* (pp. 195-213). Deventer, The Netherlands: Van Loghum Slaterus.
- MOLENAAR, I. W. (1983b). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, *48*, 49-72.
- MOLENAAR, I. W., & SYTSMA, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tydschrift Voor Onderwijsresearch*, *5*, 257-268.
- RASCH, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.
- ROSKAM, E. E., VAN DEN WOLLENBERG, A. L., & JANSEN, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement*, *10*, 265-277.
- STOKMAN, F. N. (1977). *Roll calls and sponsorships: A methodological analysis of third world group formation in the United Nations*. Leiden, The Netherlands: Sythof.
- STOKMAN, F., & VAN SCHUUR, W. (1980). Basic scaling. *Quality & Quantity*, *14*, 5-30.
- SYTSMA, K. (1984). Useful nonparametric scaling. *Psychologische Beiträge*, *26*, 423-437.
- SYTSMA, K. (1986). Another note on the usefulness of Mokken scaling. *Psychologische Beiträge*, *28*, 425-432.
- SYTSMA, K., & MOLENAAR, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*, 79-97.

(Manuscript received December 8, 1987;  
revision accepted for publication May 20, 1988.)