KM 44 (1993)
pg. 59-73

A PRACTICAL COMPARISON BETWEEN THE WEIGHTED AND THE UNWEIGHTED SCALABILITY COEFFICIENTS OF THE MOKKEN MODEL

B.T. Hemker & K. Sijtsma [1]

ABSTRACT

The Mokken model is a nonparametric item response model for ordering persons on a scale. To check if the model is valid, Loevinger's H coefficient can be used. In the case of dichotomous items, only one definition of H exists. However, for polytomous items there are two definitions, one an unweighted (H^u; Molenaar, 1982) and the other a weighted (H^w; Molenaar, 1991) H coefficient. In this paper, a practical comparison between H^u and H^w is carried out. It is found that with polytomous items, H^w results in higher values than H^u. Another finding is that H^w is not as sensitive to the number of ordered answer categories as H^u. Together with the advantages found in a theoretical study (Molenaar, 1991), these results lead to the conclusion that for standard use the weighted H coefficient should be preferred.

[1] Both authors are affiliated with Werkgroep Methoden, FSW, Rijksuniversiteit Utrecht, Postbus 80.140, 3508 TC Utrecht, tel. 030-532012.

1. INTRODUCTION

The Mokken (1971) approach to scaling entails two nonparametric item response models. These two models express the probability that a person gives a positive response to an item as a function of the person's latent trait value and the properties of the item. This function is known as the item characteristic curve (ICC). The Mokken approach is nonparametric for two reasons: the ICCs are not parametrically defined and for the purpose of parameter estimation the distribution of the latent trait need not be specified. The Mokken models thus are based on weaker assumptions than most parametric (e.g., Lord, 1980) item response models. Compared with these parametric models, the Mokken models more often fit the data. However, the Mokken approach only yields ordinal measurements, whereas the parametric models provide interval measurements.

The Mokken approach consists of two unidimensional models: the model of monotone homogeneity (MH) and the model of double monotonicity (DM). The first model requires nondecreasing ICCs. If this model fits the data, it provides an ordering of persons on a scale. The second model is a special case of the first model. In addition to nondecreasing ICCs, it requires that ICCs do not intersect. These requirements lead to an ordering of both persons and items. For many applications only the ordering of persons is needed. As a result, the first model usually receives more attention. This paper is also concerned with the MH model.

Loevinger's (1948) H coefficient is used to check if persons can be ordered accurately. For the determination of H, it is necessary to know which score patterns are error patterns in the sense of the Guttman (1944) model. In the case of dichotomous items, an error pattern for a pair of items is a score pattern with a positive or correct answer given to the most difficult of the two items and a negative or incorrect answer given to the easiest item. For a pair of items, say items i and j, the number of persons that have an error pattern is denoted by F_{ij}. Given the observed marginal distributions of a cross table containing the bivariate frequencies of scores on items i and j, the expected value of F_{ij} under the null hypothesis that the item scores are marginally independent, E_{ij}, is also determined. For an item pair, Mokken (1971) defined the scalability coefficient for two items as $H_{ij} = 1\text{-}F_{ij}/E_{ij}$.

A necessary condition for the MH model is that $0 \leq H_{ij} \leq 1$ for all item pairs. The maximum, $H_{ij} = 1$, corresponds with data on items i and j without observed error

patterns. The value $H_{ij} = 0$ corresponds with marginal independence: $F_{ij} = E_{ij}$. A score pattern without errors is called a perfect Guttman pattern. H_{ij} can be extended to coefficient H_i, which indicates whether item i is scalable in accordance with the MH model given the other items used, and to coefficient H, which is the overall scalability coefficient for k items. These latter two coefficients also range from 0 to 1, given that the MH model holds. Thus, nonnegative values are a necessary condition for this model.

Because a positive H value is not a sufficient condition for the MH model, and because low positive H values do not lead to useful scales, Mokken (1971) suggests the lower bound of $H = .30$ for practical use. Below this value the scale does not allow an accurate ordering of persons. For the interpretation of the other values of the overall scalability coefficient H, Mokken (1971) suggests the following rules of thumb:

$.30 \leq H < .40$: items form a weak scale;

$.40 \leq H < .50$: items form a medium scale;

$.50 \leq H \leq 1.00$: items form a strong scale.

The stronger the scale, the more accurately persons can be ordered. These rules of thumb are intended for practical purposes. In this study, all nonnegative H values are accepted for the purpose of comparison of different scalability coefficients.

The original Mokken model can only be used for the analysis of dichotomous items. The model was generalized to polytomous data by Molenaar (1982, 1986). This generalization is known as the polytomous Mokken model and can be applied to test items with more than two ordered response categories by means of the computer program MSP (Debets & Brouwer, 1989; Sijtsma, Debets & Molenaar, 1990).

In the polytomous Mokken approach, an adjusted scaling coefficient H is used to check if persons can be ordered accurately. However, there are two problems with implementing H in the polytomous case. The first problem is that there are two possibilities of defining a scalability coefficient. One definition implies a weighted coefficient (Molenaar, 1991), H^w, in which the weights express the degree to which an error pattern differs from a perfect Guttman pattern for polytomous items. The other definition implies an unweighted H coefficient (Molenaar, 1982), H^u, for which it is assumed that all error patterns are equally likely. Usually, the two coefficients yield different values using the same data. In this study, we provide some empirical evidence in favour of H^w.

The second problem is that it is uncertain whether Mokken's rules of thumb for the interpretation of H values can be adopted in the polytomous case. A study by Van den Berg, Sijtsma and Feij (1990), in which adjacent ordered answer categories were combined and the resulting data were reanalyzed, suggests that in the polytomous case H^u is somewhat lower than in the dichotomous case. The effect of a posteriori combining answer categories on H^w was not studied; this will be done in the present study. Before the practical comparison between H^u and H^w, however, we discuss both coefficients theoretically.

2. THEORETICAL COMPARISON OF H^u AND H^w

For the calculation of the H values for dichotomous items, the overall order of the item difficulties is needed. In the polytomous case, the item difficulty ordering is replaced by the item step difficulty ordering.

Say, an item has $m+1$ ordered answer categories and item scores $X_i = 0,1,...,m$. Response categories are indexed g and h. Each $m+1$ category item is assumed to be based on m hypothetical dichotomous item steps. An item step is the imaginary threshold between two adjacent ordered response categories.

As an example, imagine a positively worded attitude item having three ordered response categories, for example, disagree, neutral and agree. It is assumed that the subject first ascertains whether he or she agrees enough with the statement to take the first item step. If not, the first item step score equals 0, and the item score also equals 0. If the answer is affirmative, the item step score equals 1, and the subject has to ascertain whether the second item step can be taken. If not, the second item step score equals 0 and the item score equals 1. If the answer is affirmative, the second item step score equals 1, and the item score equals 2.

Note that within one item item steps are dependent. The score on step g ($g = 1,2,...,m$) of item i equals $X_{ig} = 1$ if $X_i \geq g$, and $X_{ig} = 0$ otherwise. According to this definition it is impossible for one person to have $X_{ig} = 0$ followed by $X_{i,g+1} = 1$ on step $g+1$.

In parametric item response models for polytomous items (e.g., Andrich, 1978; Masters, 1982), the item step difficulty is a latent parameter that is estimated from

observed data. In nonparametric item response models, this latent parameter cannot be estimated because the ICCs on item step level are, by definition, not parametrically defined. Therefore, the latent difficulty parameters from the parametric models are replaced by item step popularities, π_{ig}, which are the proportions of the population with $X_{ig}=1$. The proportion π_{ig} can be estimated by dividing the number of persons with $X_{ig}=1$ by the total number of persons in the sample. The order of the latent item step difficulties is assumed to be the reverse of the order of the item step popularities. Therefore, the smaller π_{ig}, the more difficult the item step. Because of the mutual dependence of item step scores within an item, by necessity $\pi_{i1} \geq \pi_{i2} \geq ... \geq \pi_{im}$, for all items i.

Using the joint order of the 2m item step popularities from two items, each with $m+1$ ordered answer categories, the error patterns on these two items can be identified. If a less popular item step of one item is passed while a more popular item step of the other item is failed, then the corresponding item score pattern (X_i,X_j) is defined as an error pattern.

As an illustration, Table 1 shows a cross table containing observed and expected bivariate frequencies for two items, i and j, each with three ordered answer categories. The popularities of the item steps can be derived from the marginal frequencies. These popularities, based on the cumulative frequencies, displayed in the last row and the last column, show that $\pi_{i1} > \pi_{j1} > \pi_{i2} > \pi_{j2}$ (a strict joint ordering). This means that item score pattern $(X_i=0,X_j=1)$ is an error pattern because the more popular item step 1 of item i was failed while the less popular step 1 of item j was passed. With two trichotomously scored items there are four different error patterns: in Table 1, in addition to the (0,1) pattern, the other three error patterns are (0,2), (1,2) and (2,0). In Table 1, the frequencies of the admissible patterns are underlined.

The number of observed error patterns for the polytomous item pair {i,j} in the sample can be written as F^u_{ij}, with superscript u denoting that errors are unweighted (that is, equally weighted). The difference with F_{ij} in the dichotomous model is that the error count is based on m^2 different error patterns rather than 1 (Van den Berg, et al. 1990). Note that $2m+1$ score patterns of the $(m+1)^2$ possible score patterns are admissible patterns. This can easily be checked in Table 1. E^u_{ij} is the expected value of F^u_{ij} given marginal independence of the item scores. The

Table 1. A cross table containing observed and expected bivariate frequencies for two three-category items

	$x_j=0$	$x_j=1$	$x_j=2$	n_i	cum
$x_i=0$	10 (8)	9 (6)	1 (6)	20	100
$x_i=1$	5 (12)	21 (9)	4 (9)	30	80
$x_i=2$	15 (20)	10 (15)	25 (15)	50	50
n_j	40	30	30	100	
cum	100	60	30		

unweighted H-coefficient is defined as $H^u_{ij} = 1\text{-}F^u_{ij}/E^u_{ij}$. For the example given in Table 1, $H^u_{ij} = .29$ ($H^u_{ij} = 1\text{-}(9+1+4+15)/(6+6+9+20) = 1\text{-}29/41 = .29$).

For the weighted H coefficient, different error patterns are weighted differently because it is assumed that some error patterns are more deviant than others. For example, in the case of the two trichotomous items in Table 1 the error pattern (0,2) is more deviant than the error pattern (0,1). The latter pattern is the product of one error. The former pattern, however, is the product of three errors: step 1 of item j is passed while the more popular step 1 of item i is failed; in addition, step 2 of item j is passed while the more popular steps 1 and 2 of item i are failed. This information is taken into account by assigning different weights to the error patterns. Let the weight $w_{xi,xj}$ denote the number of less popular item steps that are passed while more popular steps are failed to arrive at item score pattern (x_i,x_j). The weight of a pattern is defined as the number of errors that are made to obtain this pattern. Using this definition it can easily be seen that error pattern (0,2) has a weight equal to three whereas the other error patterns all have weights equal to one. By necessity, an admissible item score pattern has a weight equal to zero. See Molenaar

(1991) for more extensive examples of how to calculate weights.

If $n(x_i,x_j)$ stands for the number of times item score pattern (x_i,x_j) is observed, then F^w_{ij} can be written as $\Sigma_{xi}\Sigma_{xj}w_{xi,xj}n(x_i,x_j)$. Further, E^w_{ij} can be written as $\Sigma_{xi}\Sigma_{xj}w_{xi,xj}\epsilon n(x_i,x_j)$, where $\epsilon n(x_i,x_j)$ denotes the expected value of $n(x_i,x_j)$ under the null hypothesis of marginal independence of the item scores. A weighted H coefficient for an item pair {i,j} can be defined as $H^w_{ij} = 1-F^w_{ij}/E^w_{ij}$. For the example given in Table 1, $H^w_{ij} = .42$ ($H^w_{ij} = 1-(9+3+4+15)/(6+18+9+20) = 1-31/53 = .42$).

Just like in the dichotomous case, it is easy to extend the unweighted and weighted H coefficients for item pairs to an unweighted and a weighted H coefficient for an item (H^u_i and H^w_i, respectively) and for sets of k items (H^u and H^w, respectively). In the case of only two answer categories, $H^w_{ij} = H^u_{ij}$ because there is only one error pattern. This result also holds for H^u_i and H^w_i, and the overall scaling coefficients, H^u and H^w.

H^w has some theoretical advantages in comparison to H^u (Molenaar, 1991). If two steps of different items i and j have the same popularity ($\pi_{ig} = \pi_{jh}$, $i \neq j$), this leads to ambiguity in defining the error cell for this combination of item steps. However, the H^w_{ij} value is the same for both choices concerning the definition of the error pattern whereas the H^u_{ij} value is different. Another benefit of H^w is that, independent of the number of ordered answer categories, $H^w_{ij} = \rho_{ij}/\rho_{ij(max)}$, where ρ_{ij} is the correlation between the scores on items i and j and $\rho_{ij(max)}$ is the maximum correlation, given the marginals. For H^u_{ij}, this formula is only valid in the dichotomous case. A third advantage of H^w is that with sample size n, $E^w_{ij} = n^*cov_{ij(max)}$, where $cov_{ij(max)}$ is the maximum covariance, given the marginals. This makes it easier to calculate E^w_{ij} than E^u_{ij}.

3. PRACTICAL COMPARISON OF H^u AND H^w

3.1 Procedure

For a definite choice between H^u and H^w, their usefulness when applied to real data is investigated. Molenaar (1991) suggests that H^w often has the largest value for the same data. In this paper, this suggestion is studied using four empirical data sets. Not only the original versions of the data sets are studied but special attention is given to

the effect of combining adjacent ordered answer categories on the scalability coefficients. This is done to find out whether Mokken's rules of thumb for practical use of scalability coefficients can be applied to H^w in the polytomous case. For two datasets, different ways of trichotomization and dichotomization are studied in relation to the Hs.

Furthermore, the use of either H^u or H^w may lead to different results when items are selected for one or more scales by means of the search procedure proposed by Mokken (1971), and implemented in the computer program MSP (Debets & Brouwer, 1989). The search procedure is a stepwise bottom-up item selection procedure that only admits items with scalability values $H_i \geq c$ $(c > 0)$ to a scale. This results in a scale with an overall scalability value $H \geq c$. In this study, the search procedure is executed for H^u and H^w, respectively, for one dataset containing four subscales, using the $c = .30$ lower bound.

Finally, attention is given to the test statistic Delta Star (Mokken, 1971). This statistic is asymptotically standard normally distributed under the null-hypothesis that H is equal to zero. The statistic is used to test the null-hypothesis against the alternative that H is positive. Delta Star is calculated for the weighted and the unweighted overall H coefficients for all versions of all datasets.

All calculations are carried out with an expirimental version of the computer program MSP that included H^w. In 1993 a new version of MSP will become publicly available from iec ProGAMMA.

3.2 Data

The four data sets used in this study are called the SBL data, the Trust data, the PTP data and the Verweij data. The SBL data (Van den Berg et al., 1990) contain the scores from 441 subjects who completed the SBL-s (Spanningsbehoeftelijst-selectieversie; Van den Berg & Feij, 1988). The SBL-s is used for personnel selection. It consists of four subscales, in this paper denoted I, II, III and IV, containing 13, 20, 8 and 11 items, respectively. Each item has 7 ordered answer categories and is scored $X_i = 0,1,...,6$. Four different versions of this data set are studied: the original version, and three versions based on combining adjacent answer categories that results in alternative frequency distributions for each item (Van den Berg et al., 1990). One of these versions has five answer categories and results from combining scores 0 and 1,

and 5 and 6, respectively. Another version has three answer categories and results from combining scores 2, 3 and 4 in addition to 0 and 1, and 5 and 6, respectively. The last version represents a dichotomization of the original data by combining scores 0, 1, 2 and 3, and 4, 5 and 6, respectively. The search procedure is executed for each data set separately.

The Trust data resulted from an investigation of the attitudes of delegates of Dutch political parties (Middel & van Schuur, 1981; Sijtsma et al., 1990) with respect to their trust in people from different countries. This data set contains scores from 806 respondents on 13 four-category items. Besides the original version, two versions based on combining adjacent categories are investigated. The three-category version results from combining the two middle categories. The dichotomized version results from combining the first two categories and the last two categories, respectively.

The PTP data contain the responses of 480 subjects on the 41 items of the PTP'85 (vragenlijst voor Persoonlijk Tijdperspectief; Witjas & Oomen, 1985) that are used in the improved version of the PTP'85, the PTP'90 (Koolhaas, Sijtsma & Witjas, 1992). The PTP is a questionnaire concerning time management and subjective experience of time. The PTP data based on the PTP'90 contain eight subscales, in this study labeled A, B,..., G and J, each containing 4, 5 or 6 items. The labels I and H are skipped to avoid confusion with scale I from the SBL data and with the H coefficient, respectively. The scalability values of the subscales are studied separately. Each item is scored $X_i = 0,1,...,4$. The versions of the data that are investigated are the original version based on five answer categories, two versions based on three answer categories (one version with scores 0 and 1, and 3 and 4 combined, respectively, and an alternative version with scores 1, 2 and 3 combined) and two versions with two categories (one version with scores 0, 1 and 2, and 3 and 4 combined, respectively, and an alternative version with scores 0 and 1, and 2, 3 and 4 combined, respectively). The alternative versions are studied to find out if different meaningful ways of combining adjacent categories result in different conclusions about H^u and H^w.

The Verweij data resulted from 425 children who responded to 10 transitivity tasks (Verweij, Koops & Sijtsma, 1992; Sijtsma & Verweij, 1992). These 10 items are considered as one scale. The Verweij data are based on three ordered anwer categories ($X_i = 0,1,2$). The data are dichotomized in two different ways. The first dichotomisation (1), in which score 0 is combined with score 1, leads to a set of low

p-values for the items. The second dichotomization (2), in which score 1 is combined with score 2, yields a set of high p-values. The results of both dichotomizations are given.

3.3 Results

The results concerning the overall H^u and H^w values are given in Table 2. All (sub)scales have a H^w larger than or equal to H^u. For six of the nine original data sets with a H^u value that does not exceed Mokken's lower bound of .30, this lower bound is exceeded by H^w. In the dichotomized versions of the data, H^u and H^w are equal by necessity.

In general, H^u is larger if more answer categories are combined, which is in agreement with the results found by Van den Berg et al. (1990). One exception to this rule is provided by the second dichotomization of the Verweij data. The large difference between the two values of H for the two different dichotomizations is due to the completely different distributions of the item scores in both cases.

H^w does not seem to be very sensitive to the number of ordered answer categories. In eight cases, H^w does not change more than .03 across the different versions of the same data sets. This implies that, in practice, Mokken's rules of thumb can be applied to the analysis of polytomous items if H^w is used, regardless of the number of ordered answer categories.

The scalability values of the individual items show the same trends as the overall scalability values: H^w_i is mostly larger than H^u_i, and H^u_i increases the more answer categories are combined whereas H^w_i is hardly influenced by combining answer categories.

The alternative trichotomization and dichotomization of the PTP data do not result in scalability values that are much different from the PTP scalability values in Table 2. Thus, the conclusions also hold for these alternative versions of the data.

The SBL scales that result from the application of the search procedure to each of the four versions of the SBL data are presented in Table 3. Results are provided for H^u and H^w, respectively. For each subscale, items are marked by their item number in the original SBL subscale, making it easy to see which items are included in the newly selected scales based on the different data versions.

Table 2. The values of H^u and H^w with respect to different numbers of ordered answer categories (# cat)

The SBL data

		H^u	H^w
Subscale I			
# cat	7	.21	.35
	5	.22	.34
	3	.31	.35
	2	.34	.34
Subscale II			
# cat	7	.14	.20
	5	.15	.20
	3	.20	.20
	2	.23	.23
Subscale III			
# cat	7	.20	.35
	5	.22	.36
	3	.34	.37
	2	.43	.43
Subscale IV			
# cat	7	.14	.20
	5	.15	.21
	3	.21	.21
	2	.20	.20

The TRUST data

		H^u	H^w
# cat	4	.24	.26
	3	.31	.32
	2	.23	.23

The PTP data

		H^u	H^w
Subscale A			
# cat	5	.39	.44
	3	.37	.40
	2	.38	.38
Subscale B			
# cat	5	.36	.48
	3	.39	.47
	2	.50	.50
Subscale C			
# cat	5	.34	.41
	3	.42	.46
	2	.47	.47
Subscale D			
# cat	5	.32	.39
	3	.32	.38
	2	.38	.38
Subscale E			
# cat	5	.28	.37
	3	.31	.35
	2	.37	.37
Subscale F			
# cat	5	.28	.37
	3	.31	.33
	2	.32	.32
Subscale G			
# cat	5	.27	.34
	3	.32	.34
	2	.33	.33
Subscale J			
# cat	5	.26	.34
	3	.25	.33
	2	.34	.34

The Verweij data

		H^u	H^w
# cat	3	.36	.41
	2(1)	.76	.76
	2(2)	.20	.20

The results show that more items are selected using H^w compared to H^u. This result is found for each subscale and each version of the data, except for the dichotomized version where by necessity the same results are found for both scalabilty coefficients. Using H^w, the scales found for the four different versions of the SBL data show more resemblance than the scales found with H^u. For subscale III, in particular, using H^w the same scale is found for each version of the data. Note, however, that the resemblance between the different scales that are found for subscale II is very low. When H^u is used, the number of items included in the scale increases with increasing number of combined categories.

Within the same version of a subscale, the Delta Stars used for testing hypotheses about H^u and H^w are very similar. This test statistic decreases with decreasing number of answer categories. However, all Delta Star values are very large which is mainly due to the large sample sizes. Thus, the differences between these values have no practical consequences.

4. DISCUSSION

One important result of the practical comparison of the weighted and the unweighted H coefficients is that H^w has larger values than H^u, confirming the far more limited findings of Molenaar (1991). Consequently, the lower bound of .30 is more often exceeded by H^w than by H^u. Thus, more scales are accepted as Mokken scales and more items are admitted to a scale.

The difference between H^u and H^w is due to H^w using more information than H^u. The additional information relates to the degree of deviance of observed error patterns. Because error patterns with a large degree of deviance are more in conflict with the model than error patterns with a small degree of deviance, a coefficient using information about the degree of deviance may be preferred to one that treats all error patterns as equally deviant. Because it uses more information, H^w can better detect whether persons can be ordered than H^u. The empirical result that H^w usually leads to higher values than H^u can be considered to be a practical advantage.

A second important result from this study is that H^w is less sensitive to the number of combined ordered answer categories than H^u. Because of this sensitivity,

Table 3. SBL scales found by means of the search procedure using H^u and H^w, respectively (lower bound .30, $\alpha = .5$)

I:	13 items	
	unweighted	*weighted*
7 cat. scale:	5,7,11	1,2,3,5,6,7,8,9,10,11,13
5 cat. scale:	5,8,11	1,2,3,5,7,8,9,10,11,13
3 cat. scale:	2,3,4,5,7,8,9,11,12,13	1,2,3,5,7,8,9,10,11,12,13
2 cat. scale:	1,2,3,5,7,8,9,10,11,13	1,2,3,5,7,8,9,10,11,13
II:	20 items	
	unweighted	*weighted*
7 cat. scale:	no scale	3,15,16,17
5 cat. scale:	no scale	3,16,20
3 cat. scale:	9,14	2,3,4,15
2 cat. scale:	5,11,13,15,16,17,20	5,11,13,15,16,17,20
III:	8 items	
	unweighted	*weighted*
7 cat. scale:	5,7,8	1,2,3,4,5,7,8
5 cat. scale:	5,7,8	1,2,3,4,5,7,8
3 cat. scale:	1,2,3,4,5,7,8	1,2,3,4,5,7,8
2 cat. scale:	1,2,3,4,5,7,8	1,2,3,4,5,7,8
IV:	11 items	
	unweighted	*weighted*
7 cat. scale:	no scale	3,7,10,11
5 cat. scale:	7,10	3,7,10,11
3 cat. scale:	3,7,10	3,7,10,11
2 cat. scale:	2,3,7,10,11	2,3,7,10,11

Van den Berg et al. (1990) recommend to adapt Mokken's rules of thumb, originally proposed for dichotomous items, for use of H^u with polytomous items. The practical comparison between H^u and H^w shows that this is probably not necessary if H^w is used because H^w values are very similar for polytomous data and dichotomous data.

If H^w is used researchers may be less tempted to combine ordered answer categories to achieve better results because results are not substantially improved by this artificial manipulation. If the original data do not give satisfying results, results can not be improved by combining adjacent answer categories. However, if data based on combined categories lead to results much different from the results based on the original data (see, for example, the verweij data), this may be due to a change of the meaning of the scale (Sijtsma & Verweij, 1992).

The major practical advantages of H^w found in this study are that H^w results in higher scalability values by using more relevant information from the data, that more items are selected into one scale and that, for practical use, H^w may allow the application of Mokken's rules of thumb to the polytomous case regardless of the number of ordered response categories. Together with the theoretical advantages discussed by Molenaar (1991), these results lead to the recommendation to use H^w as the standard scalability coefficient for the MH model. H^w will be set default in the next version of MSP.

ontvangen 22-3-1993
geaccepteerd 17-8-1993

References

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

Berg, P.T. van den, & Feij, J.A. (1988). De ontwikkeling van een selectieversie van de Spanningsbehoeftelijst. *Nederlands Tijdschrift voor de Psychologie*, *43*, 328-334.

Berg, P.T. van den, Sijtsma, K., & Feij, J.A. (1990). De invloed van het samenvoegen van geordende antwoordcategorieën op de H-coëfficiënt van het Mokken-model. *Kwantitatieve Methoden*, *11*, *nr. 33*, 41-56.

Debets, P., & Brouwer, E. (1989). *User's manual MSP. Version 1.5*. Groningen: iec ProGAMMA, Rijksuniversiteit Groningen.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 255-282.

Koolhaas, M.J., Sijtsma, K., & Witjas, R. (1992). Tijdperspectieven in time management trainingen: enkele psychometrische aspecten van een vragenlijst. *Gedrag en Organisatie*, *5*, 94-105.

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

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

Middel, B.P., & Schuur, W.H. van (1981). Dutch party delegates. *Acta Politica, 16,* 241-263.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* New York/Berlin: De Gruyter.

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

Molenaar, I.W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën. In G.F. Pikkemaat & J.J.A. Moors (red.), *Liber Amicorum Jaap Muilwijk.* Groningen: Econometrisch Instituut.

Molenaar, I.W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12, nr. 37,* 97-117.

Sijtsma, K., Debets, P., & Molenaar, I.W. (1990). Mokken scale analysis for polytomous items: theory, a computer program and an empirical application. *Quality & Quantity, 24,* 173-188.

Sijtsma, K., & Verweij, A.C. (1992). Mokken scale analysis: theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education, 5,* 355-373.

Verweij, A.C., Koops, W., & Sijtsma, K. (1992). De constructie van een ontwikkelingspsychologische schaal voor transitiviteit. *Nederlands Tijdschrift voor de Psychologie, 47,* 186-194.

Witjas, R., & Oomen, P.P.M. (1985). Time management, life management. *Training en Opleiding,* 6, 10-16.