Quality & Quantity 21: 393–408 (1987)

Using Mokken scale analysis to develop unidimensional scales

Do the six abortion items in the NORC GSS form one or two scales?

MICHAEL GILLESPIE,* ELISABETH M. TENVERGERT & JOHANNES KINGMA

Department of Sociology, University of Alberta, Edmonton, Alberta, Canada, T6G 2E9
(**author for correspondence and offprints*)

Abstract. This paper describes Mokken scale analysis as a method for assessing the unidimensionality of a set of items. As a nonparametric stochastic version of Guttman scale analysis, the Mokken model provides a useful starting point in scale construction since it does not impose severe restrictions on the functional form of the item trace lines. It requires only that the item trace lines are monotonically increasing and that they do not cross. After describing the Mokken method, we illustrate it by analyzing six abortion items from the 1975–1984 NORC General Social Surveys. In contrast to earlier parametric analyses of these items (regular and probit factor analyses), we find that these items form a single dimension. We argue that the two-dimension solution of these earlier analyses is an artifact of the differences in the difficulty of the items.

Introduction

Much of the recent development in sociological methods focusses on the use of multiple items in the measurement of latent variables. This development follows two traditions which we term latent trait analysis and latent class analysis. Latent trait analysis treats the latent variable(s) as continuous, while latent class analysis treats the variables as discrete. Largely informed by classical test theory, most of the techniques in the former tradition concentrate on the reliability and validity of the items. In the LISREL-type analyses, the research typically attends to these issues in the course of developing structural equation models for the latent variables (Bohrnstedt, 1982; Jöreskog, 1982). A key feature of these models is that the most posit a linear relationship between the items and latent trait. The main exceptions are the log-linear models inspired by Rasch (1960), and the probit models of Muthen (1979, 1982).

With reference to the latent class models, these models constitute a less well developed alternative to the latent trait models that dominate sociology. Discussed mainly by Goodman (1975, 1978, 1979, 1984) and Clogg (1979) these models represent a class of log–linear models of multiway contingency tables. Although one can find substantive applications of these models in stratification research, their use in scale analysis is largely limited to method papers. An important technical difference between the two traditions is that most latent trait analyses require only the item variances and covariances, while latent class analyses typically require the full *n*-way table, i.e., higher order moments (Mooijaart, 1982). On the other hand both traditions are unified by the fact that both take local independence as the basis for measuring the latent variable, and in the case of structural equation models this basis extends to the principle of the proportionality of effects (Hauser, Tsai & Sewell, 1983; see note 1).

Despite the growing concern with measurement, researchers have neglected the question of unidimensionality, whether a set of items measure just one latent variable. In part this neglect may represent the view that the questions about reliability and validity subsume the question of unidimensionality and that unidimensionality is adequately handled by the exploratory and confirmatory factor analyses performed in the course of investigating the reliability and validity of the items. However, it can be argued that indices derived from classical test theory as well as a number of latent structures models and latent trait models are inadequate for assessing unidimensionality. In a review of classical test theory based methods for estimating unidimensionality of psychological measure, Hattie (1985) criticized the use of reliability measures and indices based on the results of factor analyses (e.g., percent variance explained by the first factor) to assess the unidimensionality of a set of items. Hattie pointed out that the high reliability is neither a sufficient nor necessary condition of unidimensionality. In addition, many of the indexes based on principal component analysis and exploratory factor analysis were shown to be inappropriate to answer the question about unidimensionality on a variety of grounds. These grounds include the statistical problems encountered in determining whether a correlation matrix has unit rank; whether to use component or factor analysis, determining the number of factors, the way of measuring communalities, using the appropriate measure of correlation and the possibility of "difficulty" factors in the case of dichotomous items.

Several investigators (Hattie, 1985; Lord, 1980) used a specific instance of the principle of local independence to define unidimensionality, that is, the existence of one latent attribute underlying the data. These researchers are more favorably disposed to methods that use the results from attempts to fit latent trait models in assessing unidimensionality. However, Hattie also noted some problems with the application of these models. First, most confirmatory factor analytic models assume a linear relationship between the items and the latent trait, whereas non-linear models may be more appropriate, particularly in the case of dichotomous items. Second, related to the problem of non-linearity is the equation of unidimensionality with a matrix of unit rank. Hattie argued that this is unnecessarily restrictive because a quadratic relation between the items and the latent trait implies a rank of two. Third, Hattie showed that the Rasch model is inappropriate to assess unidimensionality, because indices based on this model fail to differentiate between unidimensional and multidimensional items in simulation studies. This result may be attributed to the assumption of the Rasch models that the discriminatory power of the item traceline is constant across the set of items. Finally, one of the obvious problems of the goodness-of-fit test is that with large samples a chi-square is almost certain to be significant. Consequently, the null hypothesis of local independence, and therefore the unidimensionality of the set of items, is almost certain to be rejected. In order to judge unidimensionality a goodness-of-fit index is needed rather than a chi-square test and its associated probability of occurrence given the null hypothesis.

Based on the results of simulation studies, Hattie offered as promising candidates the sum of the absolute values of the residuals often "fitting a two- or three-parameter latent trait model" using methods developed by either Christofferson (1975), McDonald (1982), or Muthen (1978). However, it can be argued that a far more simple *nonparametric* Mokken (1971) method of scale analysis can be used to assess unidimensionality in the preliminary stages of scale constuction prior to subjecting the items to either calibrations of the scale values using *parametric* scaling methods or additional elaboration with the LISREL models. The Mokken method stems from Guttman's (1950) famous work on the criteria for determining whether a scale is unidimensional. However, the difficulty with Guttman's method is that it assumes a deterministic model from which the possibility of measurement error is purged. As a consequence, the item trace line or item characteristic curve (ICC), which represents the relation between the latent trait and the probability of a correct response on the item, is viewed as being a perfect step function. The difficulty this poses is that researchers have no solid criteria for deciding whether deviations from the scale types represent measurement error in a set of items that is otherwise unidimensional or whether the deviations indicate that the items lack unidimensionality.

In contrast, the Mokken method of scale analysis circumvents this problem by positing a stochastic relationship between the item and the latent variable (note 2). Taking this stochastic relationship as a starting point, the Mokken model provides a sound set of criteria for deciding both whether the set of items as a whole constitute a unidimensional set and whether a particular item should or should not be included. This last property allows the researcher to both extract a unidimensional set (assuming one exists) from a larger pool of items and see whether additional items can be added to an existing scale without reducing the unidimensionality of the scale. In addition, the Mokken model has sound criteria (indexes) for the test of goodness-of-fit, whereas in contrast to the Rasch model the ICC's do not need to have the same functional form in this Mokken model; only the principle of local independence may not be violated. The inappropriateness of the Rasch model for the assessment of unidimensionality is according to Hattie (1985) due to the fact that the discriminatory power of the items needs to be constant across the set of items (i.e., all ICC's must have the same functional form) as well as the lack of sound indexes of goodness-of-fit. Because the Mokken model does not possess these drawbacks, this model seems to be a reasonable alternative to the Rasch model for assessing unidimensionality.

The purpose of the present study was to investigate (by using the Mokken method) whether the six items on abortion, that have appeared in the NORC General Social Survey since its inception in 1972, form one or two scales (dimensions). These items constitute a particularly useful choice of an example for a study on unidimensionality because the majority of previous analysis of these items by means of a variety of methods suggested that *two* dimensions, rather than one, underlie these items (see e.g., Clogg & Sawyer, 1981; Muthen, 1982). However, Mooijaart (1982, p. 18) concluded that a one-dimensional structure can be postulated to underly these six items on abortion, i.e., a liberal versus a non-liberal attitude towards abortion.

In the next section an outline of the Mokken model is given, followed by a selective review of previous research about these six items on abortion. Subsequently, the Mokken method is applied to these items.

The Mokken model

The Mokken model is a stochastic elaboration of Guttman's scale analysis (Mokken & Lewis, 1982; Kingma & Reuvekamp, 1984, 1986a, b; Mokken, 1971). It is applied to dichotomous items for which one or the other response is designated as "positive" with respect to the attitude of interest. The model treats the attitude as a single latent trait on which the person's location is represented by the parameter θ and the item's location (difficulty) is represented by the parameter δ. Given a reasonably unidimensional set of items – that is, one dominated by the latent trait being measured, the person parameter (θ) can be estimated by the number of items to which a person responds positively, and the item parameter (δ) can be estimated by the proportion of people who respond positively. We typically refer to the former as the person's scale score and to the latter as the item difficulty.

The assumption of double monotony

The Mokken model specifies the relationship between the item and latent trait in terms of an item characteristic curve (or ICC). Let a sample of n subjects answer k dichotomous items ($X_i = 1$ when the person gives a positive response to item i and $X_i = 0$ otherwise). The probability of a positive response to item i is defined as $P(X_i = 1 | \theta_j)$. As the formal expression indicates, this curve represents the probability of a positive response on item i, given the respondent j's location θ on the latent trait. An important feature of the Mokken model is that unlike other latent trait models, it makes *no* assumption about the functional form of the ICC. For this reason, we refer to the Mokken model as non-parametric, 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 ICC's 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 positive response increases as θ increases. To put this more formally, for any two persons i and j, where θ_i is less than θ_j the probability of a positive response on any item in the scale is less for person i. The other requirement is that for any value of θ, 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 θ, or, put graphically, the ICC's do not intersect. Given this assumption, it becomes possible to define unambiguously the difficulty of an item as the θ of a person who responds positively to the item with a probability of 0.5.

Figure 1 graphically illustrates the properties of a Mokken scale. It contains the ICC's of four different items. Item 4 is the easiest, followed by items 3, 2, and 1 in order of increasing difficulty. The items satisfy the assumption of double monotony; the ICC for each item increases with θ, and none of the ICC's intersect. Note that the value of δ for each item is found by drawing a line from the ICC to the θ axis at the point on the ICC where the probability of a positive response is 0.5. Also note that the ordinal scale scores are defined in terms of the (unobserved) values of δ. Finally note that, within the constraint of double monotony, the functional form of the ICC's may differ.

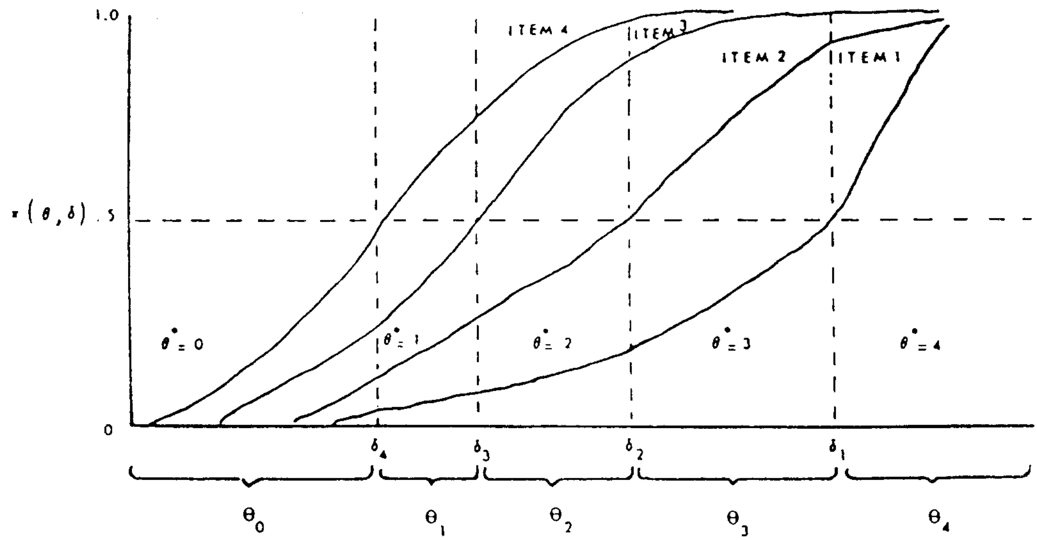

Fig. 1. An example of four item trace lines which meet the assumption of double monotony.

The value of double monotony as a criterion lies in the balance it strikes between flexibility and rigor. On the other hand, by permitting different function forms of the ICC, it avoids the problems discussed by Hattie (1985), Ten Berge (1972), and Stinchcombe (1983). On the other hand, it implies a sufficient constraint on the functional form, so that one can test the criterion with two matrices, 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 the assumption of local independence which underlies most latent trait and latent class methods. According to this assumption, item responses are conditionally independent, given the same value of θ, so that the probability of joint response for persons with the same value of θ equals the product of their marginal probabilities of response. Thus, when items 1, 2 and 3 represent decreasing levels of difficulty, the probability of a pair of positive responses will be greater for items 2 and 3, followed by items 1 and 3, and 1 and 2. Similarly, the probability of a pair of negative responses will be greatest for 1 and 2, followed by 1 and 3, and 2 and 3.

The test of double monotony, then, involves an inspection of the P- and P_0 matrices. When the rows and columns are ordered from top to bottom and from left to right according to decreasing levels of item difficulty, the probability of a pair of positive responses should increase in the (P-matrix), while the probability of a pair of negative responses should decrease in the (P_0-matrix). Deviations from this pattern can be tested with a one-sided sign test (McNemar's test for dependent proportions). Items that do not fit this pattern (i.e, that showed significant deviations from the expected pattern) are removed from the scale (Molenaar, 1982).

Coefficients of scalability

Meeting the test described above constitutes a necessary but not sufficient condition for double monotony. Further evidence is derived from three related coefficients of scalability. 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 psychometric equivalence coefficients for the relevant item pairs. The third, H, measures the homogeneity of the scale as a whole by aggregating across the coefficients for the individual items.

Due to Loevinger (1947, 1948), the coefficient for the homogeneity of an item pair essentially measures the association in the two by two table that is obtained by cross classifying the two items. Table 1 illustrates such a table. In constructing this table, we assumed that item i is more difficult than item j, and we let item i define the rows and item j define the columns.

Under Guttman's deterministic model we would expect the top right-hand 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 would expect the frequency of the error cell, $e(1, 0)$, would equal the product of the marginal frequencies divided by the sample size – i.e., $e(1, 0) = f(1, .)f(., 0)/f(., .)$. Given in (1), H_{ij}, the index of item pair homogeneity measures the proportional difference between cell frequency of the error cell expected under independence and the actual cell frequency.

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

where $e(1, 0) = f(1, .)f(., 0)/f(., .)$.

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. "1" denotes a positive response; "0" denotes a negative response.

Readers familiar with the convention of using the letters *a*, *b*, *c*, and *d* to represent the cell frequencies of a 2 × 2 table may find the following formula for H_{ij} more convenient.

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

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

The coefficient of item homogeneity, H_i, is given in (3). It simply aggregates the observed and expected frequencies used to calculate H_i for all item pairs from a set of *k* items that contain item *i*.

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

where $i \neq j$.

The coefficient of scale homogeneity, *H*, is given in (4). It aggregates the observed and expected 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 $i = j + 1$.

The coefficients of scale and item homogeneity allow the researcher to judge the scale as a whole and the scalability of individual items. Mokken (1971) has established a set of criteria for using all three coefficients to judge the homogeneity of a scale. First, all the H_{ij} should be greater than zero. Second, all the H_i and, therefore (H) should be greater than a predetermined constant (c). On the basis of his experience, Mokken gives three values of c that he proposes be used to distinguish "strong", "medium", and "weak" scales from no scale: 0.50, 0.40, and 0.30. In practice, when an item does not meet these criteria, it is eliminated from the scale. The coefficients H_i and H are then recomputed for the remaining set of items, the coefficients are checked against the criteria, and the process is repeated until a sufficiently strong Mokken scale is obtained. Computer programs developed by Niemoller et al. (1980) and Kingma and Reuvekamp (1986a, b) contain such a search procedure (note 3).

In contrast to the parametric (e.g., Rasch) models, the Mokken model is a non-parametric model, because in the latter model the functional form of the item trace line is unknown (there is only the requirement of double monotony), whereas in the Rasch models, the item trace line is a known function for the population. In the Mokken model the ordering of the items is specifically objective (an implication of the assumptions of double monotony): in any group of subjects item i is more difficult than item j (Molenaar, 1982). This property corresponds to the stronger requirement in the parametric Rasch model that the ratio of the item difficulties must be invariant across samples.

The Mokken model has proven to be very useful for scaling social attitudes, political knowledge, and political efficacy (Lippert, Schneider & Wakenhut, 1978; Mokken, 1969a, 1969b; Stokman, 1977; Stokman & Verschuur, 1980).

Data

The data came from the 1972–1982 cumulative file of the National Opinion Research Center's General Social Survey (NORC GSS). Since 1972 (with the exception of 1979 and 1981), NORC annually conducts a sample survey of persons aged 18 or over who live in the continental United States for the purpose of data dissemination and social indicator research (see for more information about the nature of these surveys, Davis & Smith, 1984).

Although NORC has obtained information on attitudes, abortion, education and church attendance on each of its annual surveys, we restricted the analysis to the 1975, 1976, 1977, 1978, 1980, 1982, 1983, 1984 surveys largely on the grounds that full probability sampling methods were not used until the 1975 survey. Combining these surveys into a single file yields a sample of 10,897 subjects exclusive of missing values.

It can be argued that in combining the different surveys into a single file, we run the risk of having our results confounded by temporal changes in public's attitudes toward abortion. We checked this possibility by means of a one-way analysis of variance of each of the six items against the year of the survey. This analysis was performed for both the full 1972–1984 cumulative file and our 1975–1984 cumulative, with, respectively, 16,432 and 10,897 cases. Of course, significant differences by year will be observed using these numbers of respondents. In both files, however, it was found that these differences account for at most one-half of one percent of the variance in any of the items, so we conclude that these changes are too small to affect the results of our scale analysis (note that Duncan, et al. (1982) draw different conclusions from their analysis of the 1973–1977 cumulative file).

We analyzed responses to questions about the respondent's approval or disapproval of legal abortion under six conditions, the so-called six abortion items. The interviewer introduced the six conditions with the statement: "Please tell me whether or not *you* think it should be possible for a pregnant woman to obtain a *legal* abortion if: (D) there is a strong chance of a serious defect in the baby; (N) if she is married and does not want any more children; (H) if the woman's own health is seriously endangered by the pregnancy; (P) if the family has a very low income and cannot afford any more children; (R) if she became pregnant as a result of rape; (S) if she is not married and does not want to marry the man (for the remainder of the paper we refer to these items as D, N, H, P, R, and S).

Results

Application of Mokken's criteria to the six abortion items yielded results which indicated that this whole set of items form a strong Mokken scale. First, the H coefficient for the whole set of six abortion items was 0.81 which, according to the cut points established by Mokken (1971), represents a very strong scale. Second, the values of scalability of the individual items, H_i's, ranged between 0.78 and 0.84 (see Table 2). These values are well above the stringent cut-point of 0.50, and thus, indicates a high scalability for each individual item of the set. Finally, inspection of the P and P_0 matrices revealed no appreciable departure from the two patterns implied by the criterion of double monotony (see Table 3).

With respect to reliability, the Kuder-Richard's formula, KR_{20}, estimate was 0.85 for these six abortion items and the item-corrected total correlations (reflecting the correlations of an item with the total score of the remaining five items) varied between 0.50 and 0.72. Applying Nunnally's (1978) criteria

Table 2. The distribution of the 6 abortion items on the Mokken scale and their P-values and coefficients H_i

Item	P-value	H_i
N	0.47	0.83
S	0.48	0.83
P	0.52	0.84
D	0.83	0.78
R	0.83	0.79
H	0.90	0.83

Table 3. The P and P_0 matrices for the six abortion items

Item	Item					
	N	S	P	D	R	H
			P matrix			
N	–	0.41	0.43	0.46	0.46	0.46
S	0.41	–	0.43	0.47	0.48	0.47
P	0.43	0.43	–	0.50	0.51	0.51
D	0.46	0.47	0.50	–	0.77	0.81
R	0.46	0.48	0.51	0.77	–	0.81
H	0.46	0.47	0.51	0.81	0.81	–
			P_0 matrix			
N	–	0.47	0.44	0.16	0.16	0.09
S	0.47	–	0.44	0.16	0.17	0.10
P	0.44	0.44	–	0.16	0.16	0.09
D	0.16	0.16	0.16	–	0.11	0.09
R	0.16	0.17	0.16	0.11	–	0.08
H	0.09	0.10	0.09	0.09	0.08	–

for a reliable test (KR_{20} > 80 and item-corrected total correlations $\geqslant 0.50$) it may be concluded that the six items may be considered a reliable scale.

Subsequently a factor analysis (principal components without iteration) of the scores of all subjects on the six abortion items revealed two principal components with eigenvalues $\geqslant 1.0$, accounting for 77 percent of the explained variance. An orthogonal Varimax rotation of the first two principal components produced the factor structure reported in Table 4. It can be seen that the items (N, S, P) with the lowest p-values have high (all > 0.87) factor loadings on the first rotated principal component and they have low factor loadings on the second rotated principal component. Furthermore, Table 4 also shows that the three easier items (D, R, H) have very low factor loadings on the first rotated principal component and high factor loadings on the second rotated principal component.

Table 4. Factor loading matrix of the first two principal components after varimax rotation for all six abortion items and the whole sample (N = 10897)

Item	Factor	
	1	2
N	0.90	0.19
S	0.88	0.22
P	0.87	0.24
D	0.25	0.82
R	0.30	0.76
H	0.09	0.86
Eigen value	3.43	1.19

In contrast to the results found with the Mokken model, these factor analytic results suggest that the abortion items represent two subscales: a measure of attitudes toward abortion for "medical" reasons (factor two) – which consists of items H, R, and D, – and a measure of attitudes toward abortion for "social" reasons (factor one) – which consists of items P, S, and N (Muthen, 1982). However, we also can interpret these results as an artifact of the different difficulty levels of the two sets of items.

Ten Berge (1972) pointed out that differences between *p*-values of items can result in artificial "difficulty factors" after factor analysis, and this is especially the case with dichotomous data as in the present study. He noted that the matrix of inter-item correlations provides evidence on the impact of differences in the difficulty level in the case of a set of items dominated by a single substantive dimension. When the items are ordered according to increasing difficulty level and the inter-item correlations in the (full) matrix are summed across the columns, the resulting row totals exhibit a circumflex pattern. Artificial difficulty factors may occur when the items with middle difficulty levels have the highest row totals, and the row totals decline as one moves toward either extreme. It was found that the correlations follow exactly this pattern. The totals for the ordering of H, R, D, P, S, and N are 3.01, 3.58, 3.62, 3.35, 3.34, and 3.00, respectively. Thus, it may be concluded that the second factor in our factor analysis captures the difference between the difficulty of the two sets of items rather than a substantive dimension.

Discussion

The question of whether the six abortion items are dominated by a single underlying dimension has been attacked directly and indirectly by researchers using a diverse array of methods: conventional scalogram analysis (Clogg & Sawyer, 1981), a variety of latent class models (Clogg & Sawyer, 1981; Mooijaart, 1982), other log–linear models (Duncan et al., 1982), a factor analytic version of multivariate probit analysis (Muthen, 1982) and, in this paper, Mokken's method of non-parametric scale analysis followed by a conventional factor analysis. Although a thorough comparison of these many methods with Mokken method lies well outside the scope of this paper, a brief discussion of these different analyses may highlight the advantages of using Mokken's method in the initial stages of scale construction (see note 4).

The difference in the analysis lies in the conclusion drawn about the number of dimensions that underlie the abortion items: one or two. This difference cuts across the distinction between latent class analysis and latent trait analysis.

With respect to latent class analysis, Mooijaart (1982) found three latent classes which he regarded as falling at different points on a single continuum: people who (tend to) respond positively to all six abortion items, people who respond positively to just the "easy" items (H, R, and D), and people who respond positively to none of the items. Similarly, the analysis reported by Duncan, Sloane and Brody (1982) also pointed to a single dimension, although with five rather than three latent classes.

In contrast, Clogg and Sawyer (1982) used an eleven "biform" model as grounds for arguing that the abortion items represent two dimensions. Ten of the classes correspond to the scale types of a conventional Guttman scale analysis. The eleventh represents a residual class of "intrinsically unscalable" individuals. Clogg and Sawyer argued that the ten scale types represent two orderings of the items. The first – H, D, R, P, S, and N – stems from the increasing level of difficulty for the six items. The second – R, H, D, N, P, and S – results from an inspection of the standardized residuals produced by a model that fits just the initial scale types plus the residual category. We note in passing, however, that the difference between these two orderings has nothing in common with the most plausible alternative to the unidimensional interpretation: one scale made up of the three "easy" items (abortion for medical reasons) and the other scale made up of the three more "difficult" items (abortion for social reasons). Thus, although further pursuit of Clogg and Sawyer's results might prove interesting, we treat these results as being beside the point of this paper.

Turning to the latent-traits methods, our Mokken scale analysis supports the conclusion that a single dimension dominates the six abortion items. However, Muthen's (1982) probit factor analysis support a two dimension conclusion: the easy items measure attitudes toward the use of abortion for medical reasons, while the more difficult items measure attitudes toward the use of abortion for social reasons. At first glance, our conventional factor analysis also seems to support the interpretation of two dimensions. However, an additional inspection with using Ten Berge's (1972) method reveals these two factors represent artificial "difficulty" factors. Therefore, we may conclude that the second factor in our factor analysis captures the difference between the difficulty of the two sets of items rather than a second substantive dimension.

The presence or absence of floor and ceiling effects in dichotomous items will determine the extent to which the relation between the item and latent trait is approximately linear. More specific, items endorsed by either an extremely low or high percentage of respondents are subject to floor and ceiling effects, respectively, and in these cases the item-trait relationship is decidedly non-linear. On the other hand, a linear approximation works well in the items endorsed by around half the respondents. Because Muthen's probit analysis posits a non-linear relationship between the item and the latent trait, it represents an improvement over conventional factor analysis. Nonetheless, we suspect that probit factor analysis fails to escape the consequences of floor and ceiling effects, because it may not capture completely the differences in the degree to which the item–trait relations depart from linearity. As Stinchcombe (1983) has pointed out, one can draw on an infinitely large family of monotonic, non-linear relationships to model item–trait relationships. In line of these arguments, by specifying a particular form of non-linear relationships for *all* items, Muthen's probit factor analysis may also produce artificial difficulty factors. In contrast, Mokken analysis is much less stringent because it makes no assumption about the functional form of the relationship between a particular item and the latent trait. It only requires that the ICC's meet the assumptions of double monotony. Therefore, the Mokken model will prove superior as a test for unidimensionality in the case of items with widely different difficulty levels.

In sum, we may conclude that the Mokken scale-analysis is a useful method for assessing the unidimensionality of a set of items. Of course, having identified a unidimensional set, further elaboration can be performed on the found scale. The researcher can see whether the set possesses additional scale properties, he/she can work on rescaling the values of resulting ordinal into a metric with interval properties. This and other tasks require the use of additional methods, but the blend of flexibility and sound statistical basis appear to make Mokken scale analysis a useful part of the initial phase of scale construction.

Notes

1. "Local independence" means that the covariance between the items will be zero when the value of the latent variable is held constant. "Proportionality of effects" refers to the condition where the covariance between an item and some criterion variable is proportional to the covariance between the item and the latent variable it measures. One can view this criterion as an extension of the principle of local independence because the proportionality of effects implies that the covariance between an item and a criterion variable will be approximately zero when the latent variable is held constant.
2. Goodman (1975) and Clogg and Sawyer (1981) also attempted to circumvent this problem by fitting a restricted latent class model to the n-way table that results from the cross-classification of the entire set of items. This model has the unique feature of partioning cases, rather than item variances into error and non-error components. It specifies $k + 2$ latent classes for a set of k items. The scale types constitute $k + 2$ classes, and the model specifies that members of these classes respond on each item with a probability of one or zero. The final class represents the random component; the members of this class are deemed "inherently unscalable". Other than the assignment of error to cases

rather than to part of the variance in the items, another major difference between the latent-class models and the Mokken (1971) method is that the former requires the complete n-way table, while the latter requires only the bivariate cross-classifications. A thorough comparison of the two approaches to unidimensionality lies beyond the scope of this paper.

3. The Mokken scale analysis is available for Apple II, II plus and IIe under CPM 2.2 operating system as well for 68000 processors under operating system CPM 68K. The program has been written in Pascal and can easily be adapted to other micro-computers and mainframes. A listing of this program and a floppy disk with the compiled version may be requested from Johannes Kingma, Department of Psychology, University of Alberta, Edmonton, Alberta, Canada, T6G 2E9.
4. The reader should note that comparison is hindered somewhat by the fact that the GSS's data analyzed differ. For example, Clogg and Sawyer (1981) analyzed just 1975 file; Duncan et al. (1982) performed analyses on the 1973–1975 cumulative file as did Muthen (1982); Mooijaart (1982) analyzed the 1972 file, while we used the 1975–1984 cumulative file for our analysis.

References

Bohrnstedt, G.W. (1982). "Measurement", pp. 69–121 in P.H. Rossi, J.D. Wright & A.B. Anderson, eds, *Handbook of Survey Research.* New York: Academic Press.

Christoffersson, A. (1975). "Factor analysis of dichotomized variables", *Psychometrika* 40: 5–32.

Clogg, C.C. (1979). "Some latent structure models for the analysis of Likert-type data". *Social Science Research* 8: 287–301.

Clogg, C.C. and Sawyer, D.O. (1981). "A comparison of alternative models for analyzing the scalability of response patterns", pp. 240–280 in S. Leinhardt, ed., *Sociological Methodology.* London: Jossey-Bass Publishers.

Davis, J.A. and Smith, T.W. (1984). *General Social Surveys, 1972–1984: Cumulative code book.* Chicago: National Opinion Research Center.

Duncan, O.D., Sloane, D.B. and Brody, C. (1982). "Latent classes inferred from response-consistency effects", pp. 19–64 in K.G. Joreskog and H. Wold, eds., *Contribution to economic analysis.* Amsterdam: North-Holland Publishing Co.

Goodman, L.A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept", *Journal of the American Statistical Association* 70: 755–768.

Goodman, L.A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis,* Cambridge, Massachusetts: Abt Books.

Goodman, L.A. (1979). "The analysis of qualitative variables using more parsimonious quasi-independence models, scaling models, and the latent structures", pp. 119–137 in R.K. Merton, J.S. Coleman and P.H. Rossi, eds., *Qualitative and Quantitative Social Research: Papers in honor of Paul F. Lazarsfeld.* New York: Free Press.

Goodman, L.A. (1984). *The analysis of cross-classified data having ordered categories.* Cambridge, Massachusetts: Harvard University Press.

Guttman, L. (1950). "The basis for scalogram analysis", pp. 60–90 in S. Stouffer, ed., *Measurement and Prediction,* Princeton, N.J.: Princeton University.

Hattie, J. (1985). "Methodology review: Assessing unidimensionality of tests and items". *Applied Psychological Measurement* 9: 139–164.

Hauser, R.M., Tsai, S.L. and Sewell, W.H. (1983). "A model of stratification with response error in social and psychological variables". *Sociology of Education* 56: 20–46.

Jöreskog, K.G. (1982), "The lisrel approach to causal model-building in the social sciences", pp. 81–99 in K.G. Jöreskog and H. Wold, eds, *Systems Under Indirect Observation.* Amsterdam: North-Holland.

Kingma, J. and Reuvekamp, J. (1984). "The construction of a development scale for seriation", *Education and Psychological Measurement* 44: 1–23.

Kingma, J. and Reuvekamp, J. (1986a). Mokken model: A Pascal program for nonparametric scaling. *Educational and Psychological Measurement* 46: 667–677.

Kingma, J. and Reuvekamp, J. (1986b). Mokken test for the robustness of nonparametric stochastic Mokken scales. *Educational and Psychological Measurement* 46: 679–685.

Lippert, E., Schneider, P. and Wakenhut, R. (1978). "Die Verwendung der Skalierungsverfahren von Mokken and Rasch zur Uberprufung and Revision von Einstellungsskallen", *Diagnostica* 24: 252–274.

Loevinger, J. (1947). "A systematic approach to construction and evaluation of tests of ability", *Psychological Monograph* 61, (No. 4, Whole no. 285).

Loevinger, J. (1948). "The technique of homogeneous tests", *Psychological Bulletin* 45: 507–529.

Lord, F.M. (1980). *Applications of Items Response Theory to Practical Testing Problems.* Hillsdale, N.J.: Erlbaum.

McDonald, R.P. (1982). "Linear versus nonlinear models in item response theory", *Applied Psychological Measurement* 6: 379–396.

Mokken, R.J. (1969a). "Dutch–American comparisons of the 'sense of political efficacy"', *Quality and 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: 425–448.

Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis.* The Hague: Mouton.

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

Molenaar, I.V. (1982). "Een tweede weging van de Mokkenschaal" (Another weighing of the Mokken scaling procedure), *Tydschrift voor Onderwijsresearch* 7: 172–181.

Mooijaart, A. (1982). "Latent structure analysis for categorical variables", pp. 77–95 in K.G. Jöreskog and H. Wold, eds., *Systems Under Indirect Observation.* Amsterdam: North-Holland.

Muthen, B. (1978). "Contributions to factor analysis of dichotomous variables", *Psychometrika* 43: 551–560.

Muthen, B. (1979). "A structural probit model with latent variables", *Journal of the American Statistical Association* 74: 807–811.

Muthen, B. (1982). "Some categorical response models with continuous latent variables", pp. 65–79 in K.G. Jöreskog and H. Wold, eds, *Systems Under Indirect Observation.* Amsterdam: North-Holland.

Niemöller, B., VanSchuur, W.H. and Stokman, F.N. (1980). *Stap User's Manual: Stochastic Cummulative Scaling*, Vol. 4. Amsterdam: University of Amsterdam.

Nunally, J.M. (1978). *Psychometric Theory*, New York: McGraw-Hill Book Co.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks paedagogiske Institut.

Stinchcombe, A.L. (1983). "Linearity in log-linear analysis", pp. 104–125 in S. Leinhardt, ed., *Sociological Methodology, 1983–1984.* London: Jossey-Bass Publishers.

Stokman, F.N. (1977). *Roll calls and sponsorships, a methodological analysis of the third world group formation in the United Nations.* Leiden: Sijthof.

Stokman, F. and Van Schuur, W. (1980). "Basic scaling", *Quality and Quantity* 14: 5–30.

Ten Berge, J. (1972). "Difficulty factors, distribution effects, and the least squares data matrix solution", *Educational and Psychological Measurement* 32: 911–920.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design.* Chicago: MESA press.