Political Analysis (2003) 11:139–163
DOI: 10.1093/pan/mpg002

Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory

Wijbrandt H. van Schuur
Department of Sociology, University of Groningen,
Grote Rozenstraat 31, 9712 TG Groningen, The Netherlands
e-mail: h.van.schuur@ppsw.rug.nl

This article introduces a model of ordinal unidimensional measurement known as Mokken scale analysis. Mokken scaling is based on principles of Item Response Theory (IRT) that originated in the Guttman scale. I compare the Mokken model with both Classical Test Theory (reliability or factor analysis) and parametric IRT models (especially with the one-parameter logistic model known as the Rasch model). Two nonparametric probabilistic versions of the Mokken model are described: the model of Monotone Homogeneity and the model of Double Monotonicity. I give procedures for dealing with both dichotomous and polytomous data, along with two scale analyses of data from the World Values Study that demonstrate the usefulness of the Mokken model.

1 Introduction

Mokken scale analysis is a combination of a measurement model and a procedure that is commonly used to assess people's abilities or attitudes. It analyzes each respondent's pattern of responses to a set of questions, or items, that are designed to be indicators of a single latent variable, i.e., the ability or attitude under study. Mokken scaling is a nonparametric probabilistic version of Guttman scaling (Guttman 1950), and it is used similarly to other techniques for data reduction that allow for the unidimensional measurement of latent variables. But it has a number of advantages over other measurement models; for example, it includes an item parameter that shows how items differ in their distribution, it is probabilistic rather than deterministic, and it can be applied in situations in which latent variables must be operationalized with only a small number of indicators.

Before presenting Mokken's scaling model and procedure in detail, I will show why its predecessor the Guttman scaling model may often be preferable to better known and more easily available measurement methods like reliability analysis and factor (or component) analysis. Next I will place Mokken scale analysis in the context of the successors to Guttman scaling, and finally introduce it in its own right. The article ends with two applications of Mokken scale analysis.

Author's note: The author would like to thank Melissa Bowerman for her editorial help, and Rob Mokken, Ivo Molenaar, and two anonymous reviewers for their comments on a previous version.

Table 1 Hypothetical data set conforming to a perfect Guttman scale

Response type	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*	*Frequency of occurrence of response pattern*
1	0	0	0	0	0	0	50
2	0	0	0	0	0	1	3
3	0	0	0	0	1	1	2
4	0	0	0	1	1	1	35
5	0	0	1	1	1	1	5
6	0	1	1	1	1	1	3
7	1	1	1	1	1	1	2
Total	2	5	10	45	47	50	100

To understand the relationship between Guttman scaling on the one hand and reliability and factor analysis on the other, consider the hypothetical data set shown in Table 1; this may look familiar to political scientists who have analyzed surveys on political participation. A hundred subjects responded to six dichotomous items. Each of seven different response patterns occurs with the frequency given in the last column. Let us imagine that V1, V2, and V3 are three indicators of party political participation (i.e., runs for office, is active in a campaign, or goes to a party meeting; 1 = yes, does; 0 = no, does not), and that V4, V5, and V6 are variables that tap voting behavior for three different types of office (i.e., president, representative, and sheriff; 1 = votes; 0 = does not vote). The first three items are rather "unpopular"—not many subjects report engaging in the activity ($p_1 = .02$, $p_2 = .05$, and $p_3 = .10$), and the last three items are noticeably more popular ($p_4 = .45$, $p_5 = .47$, and $p_6 = .50$).

This data set was constructed to form a perfectly unidimensional Guttman scale. That is, a subject who gives the positively keyed response (i.e., reports performing the activity) to a more "difficult" (i.e., unpopular) item will also give the positive response to all items that are less "difficult." The Guttman scale is therefore a "cumulative" scale: all subjects "accumulate" positive responses to the items—in this case indicators of political participation—in the same order, from the "easiest" (V6 in this example) to the most "difficult."

In this data set, the reliability (Cronbach's α) of the scale of six items is 0.84, but it could be increased to 0.86 by removing item 1, or, in fact, to 0.98 by removing all of the first three items. A principal component analysis of this data set leads to two eigenvalues larger than 1 (the first four eigenvalues are 3.36 1.71, 0.55, and 0.26). The VARIMAX rotated solution—the default in some standard statistical packages—leads to the results given in Table 2, in which the first three and the last three items load on separate components.

The results of either a principal component (or factor) analysis or a reliability analysis of this data set would have cast serious doubts on the unidimensionality of the six variables. Adherents of factor analysis might defend this outcome on grounds that all six items load highly on the first unrotated component, that one might recognize the two components as "difficulty factors," that a lower boundary of an eigenvalue of 1.00 for the substantive interpretation of a component may be too low, or that one should use tetrachoric correlations rather than product moment correlations.[1] But there are in fact good reasons to reject

[1] Embretson and Reise (2000, p. 38) note, "However, tetrachoric correlations have several disadvantages. Adjusting a whole matrix of item correlations to tetrachorics sometimes results in a singular correlation matrix, which is not

Table 2 VARIMAX rotated component loadings of data set in Table 1

	Component 1	*Component 2*
V1	0.01	0.82
V2	0.12	0.91
V3	0.28	0.79
V4	0.96	0.15
V5	0.98	0.14
V6	0.96	0.12

component analysis or reliability analysis as the appropriate way to analyze variables that form a Guttman scale.

Both component analysis and reliability analysis assume that the items can be regarded as "parallel," i.e., as having the same frequency distribution (the same mean and standard deviation). But it is precisely this assumption that is fundamentally violated in a Guttman scale: items *do* differ in their frequency distribution, as was already recognized in the 1940s by Ferguson (1941) and Carroll (1945). This difference constitutes the rationale for a Guttman scale, and it is the reason why a factor analysis of dichotomous data is difficult to interpret. The order of "difficulty" of the items often has an important theoretical interpretation that is not taken into consideration in reliability and factor analyses. If items in fact form a Guttman scale, or are expected to do so, it makes sense to analyze them with a model that takes Guttman's model assumption of cumulativity into account.

2 Guttman and His Successors: Item Response Theory

The Guttman scaling technique for analyzing empirical dichotomous data that form "quasi-scales"—i.e., that do *not* conform perfectly to Guttman's requirements—was originally introduced as a "scalogram" technique. However, it had a number of shortcomings, such as a suboptimal criterion of model fit (cf. Mokken 1971, ch. 2, for a historical overview), and was taken out of standard statistical packages. However, nothing has been introduced in its place in the computer packages, even though a number of important new developments are now available.

The new developments can be divided into two major approaches, depending on whether they regard "imperfect data"—deviations from a perfect Guttman scale—as systematic, such that they must have a substantive interpretation, or random, such that they should be treated in a probabilistic manner. Advocates of the first approach include Bart and Krus (1973), Dayton and MacReady (1980), Shye (1985), and Ganter and Wille (1999). These researchers try—each from a different perspective—to systematically represent deviations from a perfect Guttman scale by invoking one or more separate additional dimensions. Advocates of the second approach include Rasch (1960) and Mokken (1971). Their approach was first known as "Modern Test Theory" [in contrast to Reliability analysis and Component analysis, which

appropriate for factor analysis. Further, the adjustments given by the tetrachoric correlations are quite substantial for extreme p-values. Strong assumptions about normality and linearity are required to justify the adjustments. Last, tetrachoric correlations do not make full use of the data. The adjustments are based on summary statistics about the data."

are known as "Classical Test Theory" (CTT)], but it is now better known as "Item Response Theory" (or IRT for short).

In IRT models the probability of the positive response to a dichotomous item depends on both one or more subject parameters and one or more item parameters. A very popular IRT model is the one-parameter logistic model (1PL model), also known—after its inventor, the Danish statistician Georg Rasch—as the Rasch model (see Rasch 1960). In this model the probability of a positive response to an item is defined as follows (where θ_s is the parameter of subject s, δ_i is the parameter of item i, and α can be regarded as a constant for a specific scale):

$$p(x_{is} = 1 \mid \theta_s, \delta_i) = \frac{e^{\alpha(\theta_s - \delta_i)}}{1 + e^{\alpha(\theta_s - \delta_i)}}. \quad (1)$$

For $\alpha = +\infty$, the model reduces to the Guttman scale, in which the probability of a positive response approaches 1 if $\theta_s > \delta_i$, and approaches 0 if $\theta_s < \delta_i$. The Item Response Function (IRF), which depicts the probability of the positive response to item i by subjects with different subject parameters graphically, then looks like a step function (see Fig. 1). For more regular values of α (e.g., $\alpha = 1$), the IRF of a single item is a logistic (S-shaped) function, and the IRFs for different items in a scale are all parallel (see Fig. 2).

This model has the important property that subject and item parameters can be distinguished and estimated separately. The logarithm of the odds ratio of the probability of a positive response divided by the probability of a negative response to item i is $\alpha(\theta_s - \delta_i)$. The logarithm of the ratio $p(x_{is} = 1 \text{ and } x_{js} = 0)/p(x_{is} = 0 \text{ and } x_{js} = 1)$—assuming stochastic independence of the responses—is $\delta_j - \delta_i$, which is independent of subject s. Therefore, empirical estimates of this ratio lead to an estimation of the item parameters, which—if the

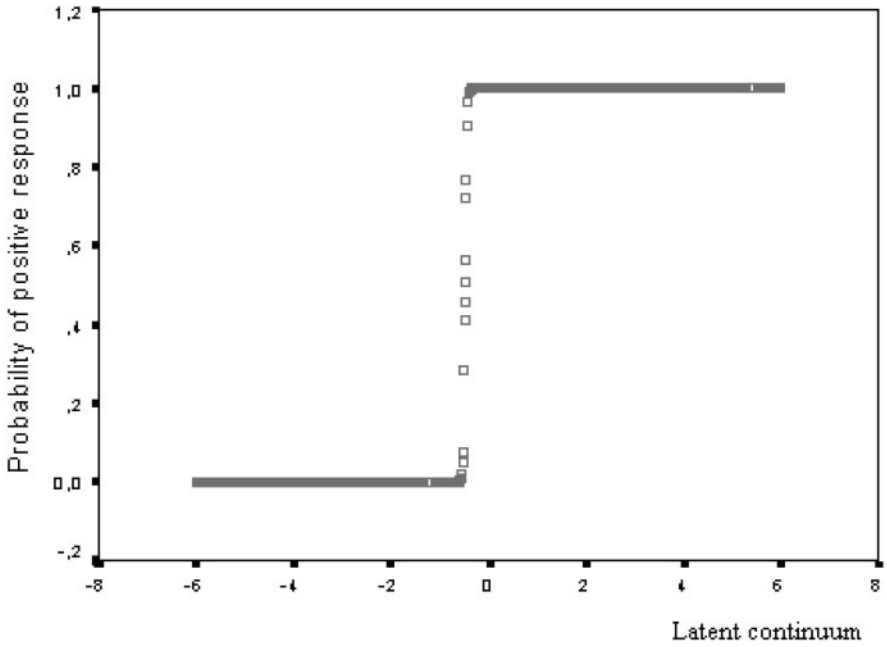

Fig. 1 Item Response Function of a Guttman item (step function).

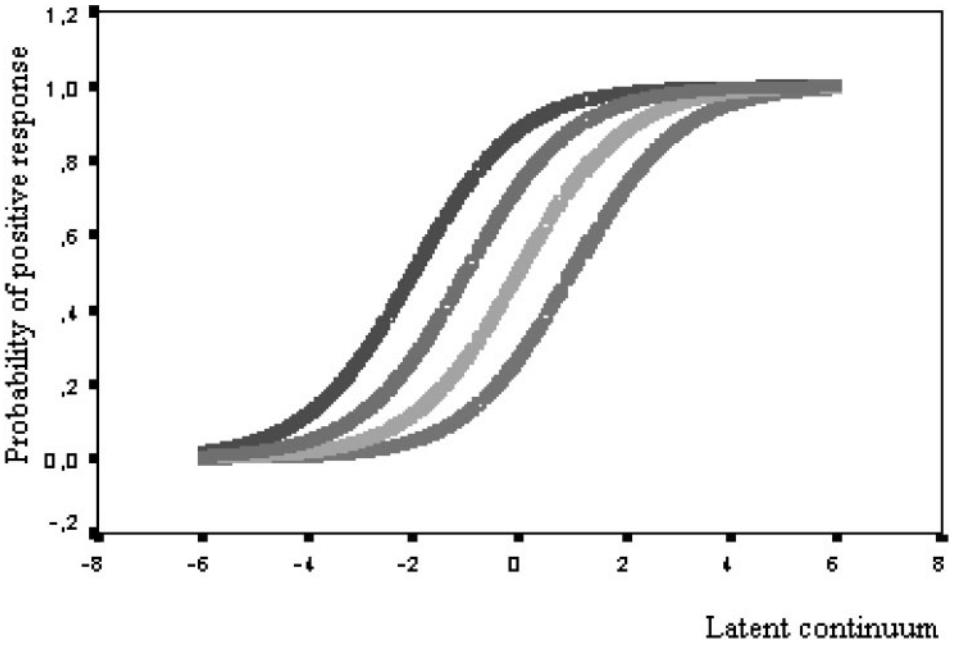

Fig. 2 Four IRFs of Double Homogeneous items that conform to the IPL model (Rasch model).

model holds—is independent of the scale values of the respondents, i.e., of characteristics specific to each respondent. Given the symmetry in the model between item and subject parameters, subject parameters can similarly be estimated independently of which specific items are used. This is equivalent to a model in which the subject and item scores are sufficient statistics. Useful introductions to the Rasch model are Andrich (1988) and Embretson and Reise (2000).

This latter property—called "specific objectivity" or "subject (or item) invariant measurement"—is crucially important in comparing subjects in different groups (e.g., across groups of subjects or over time). Andrich (1988) refers to this property as "fundamental measurement." The Rasch model has become very important in fine-tuned testing situations, especially in educational testing: it allows the equation of tests that are meant to measure the same concept at different but overlapping levels, and it allows (computerized) adaptive testing, which gives the same measurement precision with a reduced set of items.

Embretson and Reise (2000) explain the difference between CTT and IRT in 10 new measurement rules that imply the superiority of IRT over CTT. The following are two examples: (1) "Old Rule: Longer tests are more reliable than shorter tests," but "New Rule: Shorter tests can be more reliable than longer tests" and (2) "Old Rule: Comparing test scores across multiple forms is optimal when test forms are parallel," but "New Rule: Comparing test scores across multiple forms is optimal when test difficulty levels vary between persons."

Even though the Rasch model can be strongly recommended, its assumptions are strict, and the model is best applied when the number of items is rather high (e.g., greater than 20). However, in certain applications—such as omnibus-type surveys in which a large number of different topics are treated with a small number of indicators for each concept to be

measured—the Rasch model often does not fit very well. To deal with this problem, the Dutch mathematician and political scientist Robert J. Mokken (Mokken 1971; Mokken and Lewis 1982; Mokken 1997) developed two nested scaling models known jointly by his followers as "Mokken scaling" (e.g., Niemöller and van Schuur 1983). These models also conform to some of Embretson and Reise's new rules, like the two examples just mentioned. Useful introductions to the models can be found in work by Jacoby (1994, 1995); a full description is given in Sijtsma and Molenaar (2002).

One of the two models is called the model of Monotone Homogeneity (MH model), and the other the model of Double Monotonicity (DM model). The DM model has the property of Ordinal Specific Objectivity, or Invariant Item Ordering, and can therefore be interpreted as the ordinal version of the Rasch model. The IRFs in Fig. 2 conform to the DM model. The MH model can be compared to a variant of the Rasch model in which each item has a specific discrimination parameter α_i, which replaces the constant α. As this model has two item parameters (α_i and δ_i), it is known in the IRT literature as the two-parameter logistic model (2PL model), or the "Birnbaum model" after its originator (Birnbaum 1968). In the 2PL model—or, ordinally in the MH model—the IRFs of the items with their different item discrimination parameters will intersect. This implies that the order of popularity or difficulty of the items is not the same for all subjects, and item invariant measurement is not possible (see Fig. 3).

Like the Guttman scale and the Rasch model, the MH and DM models were initially developed only for dichotomous items, but by now they can also be applied to polytomous, ordered multicategory items. I will present the models for dichotomous data first, and then generalize to the polytomous case. First I present the more general of the two models, Monotone Homogeneity, and then discuss Double Monotonicity as a special case.

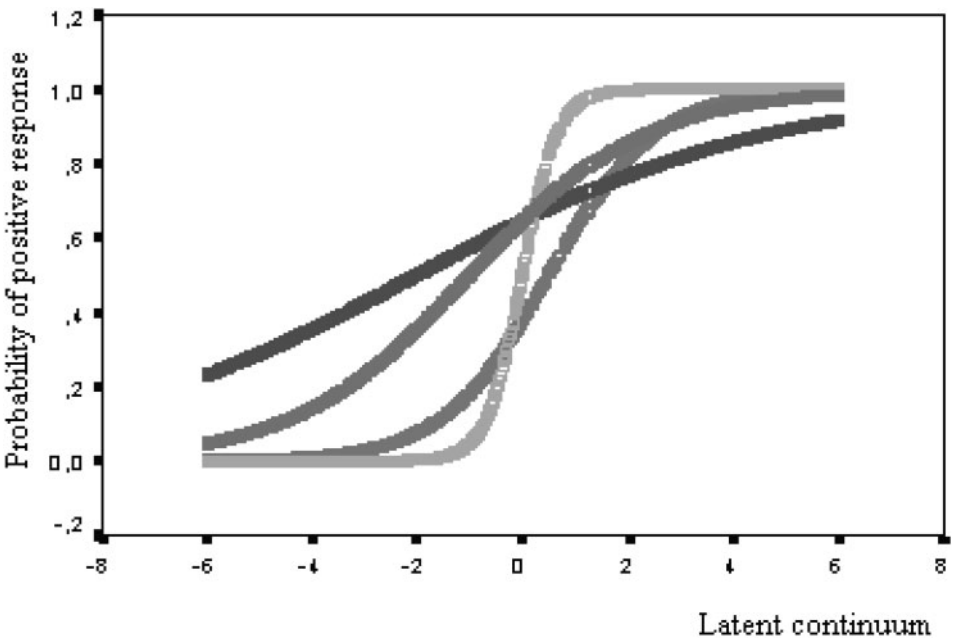

Fig. 3 Four IRFs of Monotone Homogeneous items that conform to the 2PL model.

3 The Mokken Model of Monotone Homogeneity for Dichotomous Items

The Mokken scaling model of Monotone Homogeneity makes three fundamental assumptions:

1. There is a unidimensional latent trait (e.g., an ability or an attitude) on which subjects $s \in S$ have a scale value θ_s, and on which dichotomous items $i \in I$ have a scale value δ_i. Subjects can give a positive ($x = 1$) or a negative ($x = 0$) response to each item. If the scale values of subject and item are identical, the probability of a positive response $p(x = 1 \mid \theta_s = \delta_i) = .50$. If the scale value of the subject is a lower value than that of the item, $\theta_s < \delta_i$, then $0.00 < p(x = 1) < 0.50$, and if it is higher, $\theta_s > \delta_i$, then $0.50 < p(x = 1) < 1.00$.
2. The IRF is monotonically nondecreasing. This means that the probability of a positive response to an item i increases (or at least does not decrease) with increasing subject value θ: for all items $i \in I$ and for all values $\theta_s \leq \theta_t$ we therefore assume that $p_i(\theta_s) \leq p_i(\theta_t)$. If all members of a set of items measure the same latent trait, then the ordering of the subjects by their probability of a positive response should be the same for all items. Mokken calls this "the requirement (or property) of similar ordering" of a set of items.
3. Responses by the same subject are locally stochastically independent. This means that responses to two or more items by the same subject are influenced only by θ_s, the scale value of the subject on the latent trait, and not by any other aspect of the subject or the items.

So

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{k} P(X_i = x_i \mid \theta), \tag{2}$$

in which $\mathbf{X}$ is the vector of responses to all k items, with $\mathbf{x}$ as its realization, and $P(X_i = x \mid \theta)$ denotes the conditional probability that a score of x has been obtained on item I_i. This assumption of local stochastic independence is basic to most probabilistic theories of measurement, and implies that all systematic variation in the people's responses is due only to the respondents' positions on the latent trait.

It follows from these assumptions that p_{it}—the proportion of correct answers to item i by subjects with the sum score t—is nondecreasing over increasing score groups t, where t is the unweighted sum score calculated on the basis of the remaining $n - 1$ items. This is a testable hypothesis, and it is used in testing Mokken's MH model.

If the three assumptions just discussed hold, then, as Mokken (1971) proved, all pairs of items are nonnegatively correlated for all subgroups of subjects and all subsets of items. Guttman called this requirement his First Law of Attitude. Rosenbaum (1984) proved that this result is a special case of the more general property of Conditional Association. An immediate implication, which will become relevant shortly, is that $p_{ik}(1,0)$—the probability that a subject gives the positive response to the "difficult" item i together with the negative response to the "easier" item k—is smaller than would be expected under conditional marginal independence: $p_{ik}(1,0) \leq p_i(1) \cdot p_k(0)$.

Mokken (1971, pp. 124–129) showed that the simple unweighted sum score t_s, based on the response patterns of subjects $s \in S$ (whereby subject s gets the value t as the sum of the 1- and 0-scores on the items in his response pattern), has a monotone nondecreasing regression on his scale value θ. Thus, the unweighted sum score can be used as an estimate

Table 3 Example output of the test of Monotone Homogeneity for one item

Group #	*Rest score value(s)*	*N*	*Frequencies per item value*	*Mean*	*Proportions of positive responses per item*
1	0	240	222	18	0.07
2	1	166	137	29	0.17
3	2	115	78	37	0.32
4	3–4	181	94	87	0.48
5	5	140	46	94	0.67
6	6	112	21	91	0.81
7	7	178	3	175	0.98
8	8	107	7	100	0.93

of the subject parameter θ_s, just as in reliability analysis. Other methods of calculating scores have been proposed (e.g., Schriever 1985) but will not be considered here.

3.1 *Testing Whether a Set of Items forms a Mokken Scale that Conforms to the Requirement of Monotone Homogeneity*

How do we know whether a set of items forms a Mokken scale? I will first give a simple test of Monotone Homogeneity, and then describe a full search procedure for finding a maximum set of items, from a potentially larger pool, which conform to the requirements of a Mokken scale.

The test of Monotone Homogeneity is rather simple, and can best be illustrated with an example. Given a Mokken scale with k items, we test whether the probability of a positive response to any given item increases with increasing scale values of the subjects. A subject's scale value is based on the rest scores—the subject's sum score on the remaining $k - 1$ items. The $k - 1$ items give rise to k different possible scale values $(0 - (k - 1))$, and so we can differentiate the subjects into k different groups, depending on their rest score.

My example is based on the responses of 1019 Dutch respondents in the European Values Study 1990, who answered nine questions about religious beliefs, e.g., "Do you believe in God, in Heaven, in the Devil, in Hell?" "Yes" was the positive response and "No" the negative response. The question singled out for inspection here is "Do you believe in reincarnation?" For this test I have specified that each of the (k) rest score groups should consist of at least 100 subjects. With $k = 9$ items, this requirement led me to collapse the adjacent rest score groups 3 and 4. The results are given in Table 3.

The first rest score group (group 1, with the scale score of 0 on the remaining eight items) consists of 240 subjects. Of these, only 18 gave the positive response to the ninth (reincarnation) item. So for this rest score group the proportion of subjects who gave a positive response to the ninth item was $18/240 = 0.07$. Rest score group 2, with scale

Table 4 Rest score groups 7 and 8 by item scores

	Item: score 0	*Item: score 1*	*Total*
Rest score 7	3	175	178
Rest score 8	7	100	107
Total	10	275	285

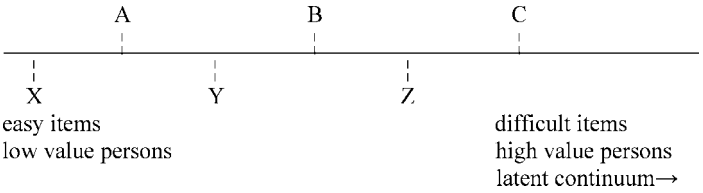

Fig. 4 Three persons (X, Y, and Z) and three items (A, B, and C) along a Guttman scale.

score 1 on the remaining eight items, consists of 166 subjects, of whom 29 gave the positive response to the reincarnation item, which leads to a proportion of $29/166 = 0.17$. As under the assumption of Monotone Homogeneity subjects in rest score group 2 are expected to have a higher value on the latent dimension than subjects in rest score group 1, they are also expected to have a higher probability of giving a positive response to the ninth item. Indeed, 0.17 is higher than 0.07. However, the pattern of increasing proportions of positive responses to the ninth (reincarnation) item for subjects with increasing scale scores on the remaining eight items is violated by the last two scale score groups, because the proportion of positive responses for group 7 (0.98) is higher than for group 8 (0.93).

Is this violation statistically significant, or is it due to sampling variation? To decide, we construct the crosstable for the two rest score groups, and test the null hypothesis that $p(x = 1)$ for group 7 is equal to $p(x = 1)$ for group 8, using the hypergeometric distribution (see Table 4).

The exact probability of this degree of violation is the probability of a cell value of 100 or less [for cell (rest score 8, item score 1)] with marginals 107, 178, 10, and 275. A good normal approximation to this hypergeometric distribution has the normal deviate $z = 2[\sqrt{(101 \times 4)} - \sqrt{(175 \times 7)}/\sqrt{286}] = 1.762$ (see Molenaar 1973, for the derivation of this calculation). With a one-sided significance level of 5%, this z-value is higher than the critical value of 1.64. The violation of Monotone Homogeneity shown by rest score groups 7 and 8 is thus too large to be ascribed to sampling error. The researcher might therefore decide to discard the item from the scale.[2]

3.2 *The Concepts of Deterministic Model Violation and Homogeneity*

We started our exposition of the Mokken scale with the test of Monotone Homogeneity, which is a simple nonparametric probabilistic test, to demonstrate the probabilistic nature of this Mokken model. But the Mokken scaling procedure is better known by its practitioners for its bottom-up hierarchical clustering procedure, which identifies a maximal subset of items that conform to the requirements of a Mokken scale. This procedure is generally described in terms of amount of violation of the deterministic Guttman scaling model, although it can be described in probabilistic terms as well. First, consider model violations in a Guttman scale, which will allow me to introduce an important test used in Mokken's scaling procedure: Loevinger's coefficient of homogeneity (Loevinger 1948).

Let us assume three items, A, B, and C, in order of difficulty, and three persons, X, Y, and Z, in order of ability (Fig. 4).

In a perfect Guttman scale, subjects will give the positive response to items represented to the left of their position (i.e., the "easier" items), and the negative response to items

[2]However, a statistically *significant* model violation can pragmatically be considered as not sufficiently substantively *relevant*, and other pragmatic considerations (e.g., an already small number of items) may lead the researcher to reconsider whether this violation is serious enough to warrant discarding the item.

Table 5 Three crosstables with the same marginal distributions, comparing an empirical distribution with two theoretically expected distributions

	Empirical (observed) distribution (hypothetical)			*Expected distribution in an ideal cumulative scale*			*Expected distribution under the assumption of stochastic independence*		
	$i=1$	$i=0$	*Total*	$i=1$	$i=0$	*Total*	$i=1$	$i=0$	*Total*
$j=1$	18	2	20	20	0	20	12	8	20
$j=0$	42	38	80	40	40	80	48	32	80
Total	60	40	100	60	40	100	60	40	100

represented to the right (i.e., the "more difficult" items). Four types of transitivity relations are possible, of which only one can violate the model[3]: if subject Z dominates item B (i.e., subject Z gives the "positive" response to item B), and item B dominates item A (i.e., item B is "more difficult" than item A), then subject Z will (i.e., is expected to, according to the model) dominate item A as well. This violation can be empirically observed: a subject who gives the positive response to a difficult item should also—but does not necessarily—give the positive response to an easier item.

The concept of model violation thus revolves around a triple of objects consisting of one subject and two items. The number of model violations in a data set is defined as the number of transitivity relations among all such triples that are violated. As both subjects and items are involved, it is possible to attribute model violations to either subjects or items. It is usually not clear from the data or the theory as to which attribution is more plausible. But it is generally in the interest of researchers to keep their (representative) samples intact, and not to draw conclusions unless they can be supported by the whole sample and generalized to a wider population. Researchers may also be interested in identifying the most prototypical indicators of a specific concept, and judge that deleting items that are not homogeneous with the rest contributes to better measurement. Researchers therefore usually attribute violations to items rather than subjects, and we will initially do the same. Later, however, model violation will also be defined on the basis of the subjects, in terms of "person fit."

Homogeneity, whether of items or subjects, is defined by relating the number of model violations observed [denoted as the number of "errors observed" or "E(obs)"] to the number of violations that can be expected under the model of stochastic independence [denoted as "E(exp)"]. This can be shown very simply with the example in Table 5.

Consider two items with popularity $p_i = .60$ and $p_j = .20$. Item j is the more difficult of the two, so in the situation of the perfectly cumulative model the cell ($j = 1, i = 0$) should be empty. In the case of stochastic independence the frequency of this cell is calculated as the product of the marginals, divided by the grand total: $(20 \times 40/100) = 8$. The observed frequency of this "error cell" is 2. Following Loevinger (1948), the homogeneity of a pair of items can be defined as $H_{ij} = 1 - E(\text{obs})/E(\text{exp})$, in which $E(\text{obs}) = 2$ and $E(\text{exp}) = 8$, or $H_{ij} = 0.75$. This coefficient of homogeneity can be rewritten as the ratio of the covariance between items X_i and X_j and their maximal obtainable value, given the marginal distribution of both items.

The homogeneity of the entire scale, H, is defined in terms of the ratio of the total sum of all errors observed versus expected, or, alternatively, as the ratio of the sum of all pairwise

[3] It is not possible to violate transitivity relations (1) among three items, or (2) among three subjects, or (3) among two subjects and one item.

covariances versus the sum of all pairwise maximal covariances (Mokken 1971):

$$H = 1 - \frac{\sum_{i=j+1}^{k} \sum_{j=1}^{k-1} E(\text{obs})_{ij}}{\sum_{i=j+1}^{k} \sum_{j=1}^{k-1} E(\text{exp})_{ij}} \quad \text{or} \quad \frac{\sum_{i=j+1}^{k} \sum_{j=1}^{k-1} \text{cov}(X_i, X_j)}{\sum_{i=j+1}^{k} \sum_{j}^{k-1} \text{cov}(X_i, X_j)_{\max}}. \tag{3}$$

Item coefficients of homogeneity, H_i for item i, are similarly defined as

$$H_i = 1 - \frac{\sum_{j=1, j\neq i}^{k} E(\text{obs})_{ij}}{\sum_{j=1, j\neq i}^{k} E(\text{exp})_{ij}} \quad \text{or} \quad \frac{\sum_{j=1, j\neq i}^{k} \text{cov}(X_i, X_j)}{\sum_{j=1, j\neq i}^{k} \text{cov}(X_i, X_j)_{\max}}. \tag{4}$$

3.3 *Statistical Test and Search Procedure*

The coefficients of homogeneity, H_{ij}, H_i, and H, play a central role in testing or constructing a Mokken scale. It is important not to conclude that the homogeneity in a specific data set is "high enough" without considering whether—given the sample size and the item difficulty—the values could have been attained by chance. To control for this, Mokken (1971, pp. 160–164) formulated a test of the null hypothesis that H_{ij} (or H_i, or H) is 0 in the population, i.e., that all item pairs are stochastically independent of one another. He showed that for the H_{ij} coefficient, as the simplest example, the statistic $Z_{ij} = (N-1)^2 \cdot (p_{ij} - p_i \cdot p_j)/[p_i(1-p_i) \cdot p_j(1-p_j)]^2$ has a standard normal distribution for a large N (i.e., a large number of subjects in the sample). Statistics Z_i and Z, which test the null hypothesis for H_i and H, respectively, are derived similarly.

As a rule of thumb, Mokken (1971) suggested that in order to accept a set of items as a Mokken scale, we require that each $H_{ij} > 0$, and that each $H_i > 0.30$. This also implies that $H > 0.30$. This rule has withstood the test of time. As the H_{ij} coefficient can be interpreted as the correlation coefficient between items i and j, divided by its maximal attainable value given the marginal frequencies of both items, the lower boundary of homogeneity of each item ensures that most correlations are substantial. In terms of the Rasch model [see Eq. (1)] this is equivalent to finding a high enough value for α that corresponds to well-discriminating items, or in more practical terms, to a high enough index of person separation (Andrich and Douglas 1977).

When items do not constitute a homogeneous set of indicators of the same latent trait, most scaling and other data reduction techniques—including reliability analysis, factor analysis, and Rasch scaling—use a "top-down" approach to find the "best" subset of indicators, first investigating the whole set and then deleting the "worst" item (in reliability analysis this is the "alpha if item deleted" procedure). In contrast, the search strategy in the Mokken scaling procedure consists of a "bottom-up" hierarchical clustering procedure, with the homogeneity coefficient serving as the clustering criterion. The procedure consists of the following steps:

a. Find the best smallest scale, i.e., that pair of items with the highest H_{ij} coefficient. This coefficient must be higher than a user-specified lower boundary (usually 0.3). If two or more item pairs are tied for the highest coefficient, select the pair with the most difficult (or "unpopular") items.
b. Find the next best item in the scale, and iterate this step. For each next best item we require the H_{ij}'s for this item, in combination with all other items already in the scale, to be positive, and the item's H_i to be higher than the user-specified lower boundary. Z_i must be larger than a user-specified criterion (usually related to an α-level of 0.05). In order to avoid "capitalizing on chance," as is possible in hierarchical clustering

procedures, the α-level is decreased each time a new item has been added to the scale. The new scale should have the highest H-value, compared to the addition of any other item. If two or more items are tied for the highest H-value, select the item with the highest H_i-value, and—if necessary to discriminate between items—the most difficult item. The search procedure stops when there are no further items that fulfill these requirements, or when all the items have been incorporated into the scale.

Sometimes an item that was selected in an earlier phase of the search procedure may, by the end of the procedure, have an H_i-value that has fallen below the user-specified boundary. Such an item may need to be removed. But in practical applications over the last 30 years this has rarely happened.

3.4 *Person Fit: Number of Guttman Errors*

A violation of the deterministic cumulative model is, as noted, most typically attributed to one or more "bad" or nonhomogeneous items. But sometimes it is useful to investigate the subjects who stand out with unlikely response patterns. Why do these subjects behave differently? We first determine which subjects are the least homogeneous with the rest of the sample by calculating and comparing the total number of simple "Guttman errors" in all the subjects' response patterns—i.e., the transitivity violations (Meijer 1994). For example, if we have five items ABCDE, ordered from easy to difficult, then subject 1, with response pattern 10101, makes three Guttman errors (violations in the pairs BC, BE, and DE), and subject 2, with response pattern 00111, has six errors (the pairs AC, AD, AE, and BC, BD, BE). Subjects' numbers of Guttman errors can be regarded as their values on a new variable (see also Meijer and Sijtsma 2001, for a recent overview of person-fit measures). The Mokken scaling program (MSP) determines the number of Guttman errors for each subject, and stores this in a file that can be used for additional analysis.

3.5 *The Model of Double Monotonicity and Accompanying Model Tests*

Each item in a set of monotone homogeneous items allows us to rank a set of persons in the same order on the latent continuum. However, Monotone Homogeneity is not sufficient to establish a uniform rank ordering of items, i.e., an ordering in which all subjects perceive the items in the same rank order of difficulty. To ensure the same ordering of items by subjects a stronger requirement is needed. This is the model of Double Monotonicity (Mokken 1971).[4] A set of monotone homogeneous items, I, satisfies the condition of Double Monotonicity with respect to a set of subjects, J, irrespective of their value on the latent continuum, if, for all pairs of items $(i, k) \in I$, it holds that if for some subject with scale value θ_0, $p_i(\theta_0) < p_k(\theta_0)$, then for all subjects $\in J$, irrespective of their value on the latent continuum, $p_i(\theta_j) \leq p_k(\theta_j)$, where item i is assumed to be the more difficult of the two items. This is the ordinal variant of Rasch's requirement of specific objectivity, or item-independent subject measurement, a model property that increases the validity of comparisons of scale scores of subjects on the same scale in different data sets.

According to the model of Double Monotonicity, the order of the manifest probabilities p_i reflects an ordering of the items according to their difficulty that is uniform across (sub)groups of persons. The more general model of Monotone Homogeneity, discussed previously, does not imply this: the item response functions of any two items may intersect

[4] Rosenbaum (1987) introduced an equivalent concept, according to which one item is uniformly more difficult than another when its item response function is higher or equal to that of the other.

because they may increase with different slopes. For two groups of subjects, one with a lower and one with a higher value than the scale value indicated by the intersection point, the manifest probabilities will indicate different orders of difficulty of the items.

There are several procedures for assessing whether a set of items conforms to the requirement of Double Monotonicity. As a first procedure, one can test whether the order of difficulty of the items is the same across subgroups of subjects (e.g., men vs. women, age groups), using the order of the marginal probabilities of the positive response for the total sample as a baseline. In the case of discrepancies from the baseline, a binomial test is carried out (Molenaar 1973) to determine whether the samples might nevertheless have come from populations in which the items have the same marginal probability. If this hypothesis must be rejected, then the model assumption of Double Monotonicity is violated for this pair of items in these subgroups of subjects.

Two other procedures, similar to the test of Monotone Homogeneity, are available to compare the order of the probability of the positive response to pairs of items for groups of subjects that are distinguished by their sum scores on the remaining $k - 2$ items: the "restscore group method" and the "restscore splitting method" (see Sijtsma and Molenaar 2002, for elaboration).

A fourth test is based on the probability of joint (non)occurrence of pairs of items in subjects' response patterns. If, according to subject j, the order of difficulty of three items, i, k, and l, is $i < k < l$ [or, $p_i(\theta_j) \leq p_k(\theta_j) \leq p_l(\theta_j)$], then, given local stochastic independence, the order of the probabilities of giving the positive response to both items in each pair of items is

$$p_{ij}(\theta_j) \leq p_{ik}(\theta_j) \leq p_{jk}(\theta_j). \tag{5}$$

If the order of the marginal popularity of the three items is the same for all subjects, then the order of probabilities shown in Eq. (5) holds for all subjects, and the following relationship between the probabilities of the positive response to both items in the item pairs should be observed:

$$p_{ij} \leq p_{ik} \leq p_{jk}. \tag{6}$$

In a square symmetrical matrix (called the $\mathrm{P}^{1,1}$ matrix) in which the items are ordered from difficult to easy, the cell elements—which contain the proportion of subjects who give the positive response to both items in the pair—are hypothesized to increase from left to right and from top to bottom. To test Double Monotonicity we can test this hypothesis.

The cell proportions of the joint negative responses (i.e., the proportion of subjects who give the negative responses to both items in a pair) in such a matrix (called the $\mathbf{P}^{0,0}$ matrix) should decrease from left to right and from top to bottom. Comparable results have been demonstrated for the $\mathbf{P}^{1,0}$ and the $\mathbf{P}^{0,1}$ matrices, in which only one of the two items in an item pair gets the positive response (Coombs and Lingoes 1978; Rosenbaum 1987).

3.6 *Deciding When to Discard an Item*

Different model tests of Monotone Homogeneity or of Double Monotonicity may single out different items as violating the model. What to do then? The best strategy is to remove one item at a time to see what effect this has on the rest score for the other items because the remaining number of violations for pairs involving these other items may be reduced by different amounts. But which item should be discarded first? As simulation studies have

shown, valuable information on an item's suitability for the scale can be gleaned from a variety of goodness-of-fit indicators, including (a) how far coefficient H_i falls below the user-specified lower boundary; (b) how many model violations there are, both in absolute terms and relative to the maximal possible number of violations; and (c) how serious the violations are in terms of z-scores that test the significance of the binomial null hypotheses. These different indicators have been combined into a single statistic called CRIT by Molenaar and Sijtsma (2000, p. 74). The item with the highest CRIT value (as a rule of thumb: CRIT $>$ 80) should be discarded first. The interpretation of this statistic is still under study.

3.7 *An Estimate of Reliability in Mokken Scale Analysis*

In the earlier discussion of a model test for Double Monotonicity, we defined a square, symmetrical $\mathbf{P}^{1,1}$ matrix of order $k \times k$, in which the rows and columns contain the items ordered from difficult to easy, and the cells contain the proportion of subjects who give the positive response to both items in the pair. If the items conform to the requirement of Double Monotonicity, cell proportions increase from both left to right and top to bottom.

The diagonal cells in this matrix are obviously empty, but they can be filled by interpolation (or extrapolation, for the first and last cell). The diagonal cells give an estimate of the probability of a positive response to the same item if it were offered for a second time (i.e., p_{ii}). This probability is used in the test–retest reliability coefficient ρ, comparable to Cronbach's α, first proposed by Mokken (1971, pp. 142–147), and improved by Sijtsma and Molenaar (1987):

$$\rho = \frac{\sum_{i=1}^{k}\left[p_{ii} - p_i^2\right] + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}[p_{ij} - p_i \cdot p_j]}{\sum_{i=1}^{k}[p_i(1-p_i)] + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}[p_{ij} - p_i \cdot p_j]}. \tag{7}$$

4 Extending the Mokken Model to Polytomous Items

So far we have discussed Mokken scale analysis as a measurement model and a procedure only as applied to dichotomous data, as in the original proposals by Guttman and Rasch. But many items that measure abilities or attitudes in survey analysis have more than two ordered categories, e.g., disagree strongly, disagree, agree, and agree strongly. These are called multicategory or polytomous items. Of course, polytomous data can generally be recoded as dichotomous data, although information is lost. But the Mokken scaling model has also been extended to polytomous items.

Table 6 shows a hypothetical data matrix of six variables with four categories each. When these categories are recoded such that (0,1 = 0) and (2,3 = 1), the data matrix is identical to the one shown in Table 1 in the introduction. The six questions might deal with voters' confidence in different national institutions, e.g., "How much confidence do you have in the Trade Unions (V1), Parliament (V2), the Press (V3), the Legal System (V4), the Police (V5), or the Church (V6)?": "none at all" (0), "not very much" (1), "quite a bit" (2), or "a great deal" (3). Respondents' responses to these questions can be used as indicators of how much confidence they have in their society as a whole, that is, the sum score of respondents' answers to the six questions can be regarded as a measurement of their general feeling of confidence.

For the data set shown in Table 6, Cronbach's α is 0.90, but it could be increased to 0.92 by removing item 1, and it could be increased to 0.97 by removing the first three items. A component analysis of this data set leads to two eigenvalues larger than 1 (the first three eigenvalues are 4.25, 1.39, and 0.22). The results of the VARIMAX rotated solution are

Table 6 Hypothetical data set of six variables with four categories each [which, after recoding (0,1 = 0, 2,3 = 1) is identical to the data set of Table 1]

Response type	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*	*Frequency of occurrence of response pattern*
1	0	0	0	0	0	0	25
2	0	0	0	0	0	1	25
3	0	0	0	0	0	2	1
4	0	0	0	0	1	2	2
5	0	0	0	1	2	2	1
6	0	0	0	1	2	3	1
7	0	0	1	2	2	3	17
8	0	0	1	2	3	3	18
9	0	1	2	2	3	3	2
10	0	1	2	3	3	3	3
11	0	2	2	3	3	3	1
12	1	2	3	3	3	3	2
13	2	2	3	3	3	3	1
14	3	3	3	3	3	3	1

similar to those of the component solution shown in Table 2, with the first two and the last three items loading on separate components.

This data set was constructed as a perfect unidimensional Guttman scale for polytomous items, comparable to our first example of a Guttman scale for dichotomous data. Again, the standard analysis techniques (reliability analysis and factor analysis) cast doubt on the unidimensional interpretation of the data. To do justice to the cumulativity of the responses, a measurement model that incorporates parameters for different values of the items is needed.

In his polytomous extension of the dichotomous Mokken model, Molenaar (1991, 1997b) solved this problem by introducing the concept of an "item step." For each item i with m categories, the latent trait is divided into m ordered areas that are separated by $m - 1$ item steps: the positions on the latent continuum that distinguish between the subjects in or below the m-th category and the subjects in or above the $(m + 1)$-th category of item i, denoted as δ_{i01}, δ_{i12}, and δ_{i23} (see Fig. 5 for an example with $m = 4$).

The $m - 1$ item steps can be regarded as $m - 1$ new dichotomous variables. For instance, the values of V_i are recoded into $(1,2,3 = 1)(0 = 0)$ for the new dichotomous variable V_{i01}, into $(2,3 = 1)(0,1 = 0)$ for the new dichotomous variable V_{i12}, and into $(3 = 1)(0,1,2 = 0)$ for the new dichotomous variable V_{i23}. The concept of "cumulativity" now is related to these new variables: the item steps, rather than the original items, are now ordered in terms of their "difficulty."

These $m - 1$ new variables are not independent of each other. For each original variable V_i, the difficulty order of the new item step variables is fixed with the following order of the item step parameters (from easy to difficult): $\delta_{i01} < \delta_{i12} < \cdots < \delta_{im-1,m}$. For each *pair* of polytomous items, however, the difficulty order of two item step variables, each from one of

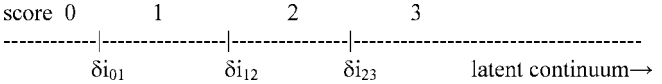

Fig. 5 Position of the three item steps of item i with four ordered categories on a latent continuum.

score: i	0	0	0	1	1	2	3
score: j	0	1	2	2	3	3	3
freq in example	20	10	10	10	10	30	10
sum score	0	1	2	3	4	5	6
		δ_{j01}	δ_{j12}	δ_{i01}	δ_{j23}	δ_{i12}	δ_{i23}

Fig. 6 Example of the position of the six item steps of two items with four ordered categories each.

the items, is *not* fixed.[5] For instance, δ_{j12} may be represented anywhere with respect to all of the $m-1$ item step variables of item i. For two original variables with three categories each, e.g., i_{01}, i_{12}, j_{01}, and j_{12}, there are six possible orders of the item steps, even though $i_{01} < i_{12}$ and $j_{01} < j_{12}$: (1) $i_{01} < i_{12} < j_{01} < j_{12}$; (2) $i_{01} < j_{12} < i_{12} < j_{01}$; (3) $i_{01} < j_{12} < j_{01} < i_{12}$; (4) $j_{01} < j_{12} < i_{01} < i_{12}$; (5) $j_{01} < i_{12} < j_{12} < i_{01}$; and (6) $j_{01} < i_{12} < i_{01} < j_{12}$. It is not the case, for instance, that all item steps of one item must precede or follow all item steps of another item.

They are distinguished by the different orders of popularity of the item values. Figure 6 shows only one of these orders (the frequencies in this example are hypothetical, and are immaterial to the number of possibilities).

The homogeneity of two (or more) variables is determined by comparing the cell frequencies of pairs of polytomous variables in their $m \times m$ crosstables. For dichotomous data, each crosstable contains one "error cell." For polytomous items, however, crosstables contain more than one "error cell," although these do not all violate the Guttman model to the same extent, as will be demonstrated below. The homogeneity of a pair of items, H_{ij}, is still defined as $1 - E(\text{obs})/E(\text{exp})$, except that now the number of model errors observed and expected are based on the summation of the different error cells in the crosstable.

Each order of the item step variables gives rise to a crosstable with a specific order of popularity of the item values. In this table, the deterministic model excludes specific combinations of responses to the two items. An $m \times m$ crosstable has m^2 cells. Each pair of original variables has $2m-2$ item step variables that divide the latent trait into $2m-1$ areas. These areas are represented by $2m-1$ cells in the crosstable. The number of error cells—i.e., combinations of values of the original variables that violate the Guttman model—in the crosstable therefore is $m^2 - 2m + 1$. So for two 4-category items, the number of error cells is $16 - 8 + 1 = 9$. Table 7 gives for each cell two numbers: first the expected cell frequency, given a perfect Guttman scale with the given marginal frequencies for the two variables i and j, and second, between brackets, the expected cell frequency if the two items are stochastically independent. The values of the marginal frequencies of each of the ordered categories are used to determine which cells in the crosstable contain no model violations if the deterministic model fits. The cell values in the whole crosstable now show the distribution of the pair of variables with the maximum possible covariation. The empirically observed frequencies in the cells that should be empty under the deterministic model are the numbers of error observed (these are not shown in Table 7).

The error cells in the crosstable differ, as mentioned earlier, in how seriously the responses they represent violate the model. This can best be seen if we regard the item step variables as new dichotomous variables for which the original Guttman model should hold. Each response to a pair of original items is now translated into a response pattern to the new

[5] In parametric IRT this freedom in the ordering of the item steps is comparable to the Graded Response Model (Samejima 1969).

Table 7 Crosstable for two items in which the cell frequencies conform to either the perfect Guttman scale, or (between brackets) the model of stochastic independence

	$j=0$	$j=1$	$j=2$	$j=3$	*Total* $I=i$
$i=0$	20 (8)	10 (4)	10 (8)	0 (20)	40
$i=1$	0 (4)	0 (2)	10 (4)	10 (10)	20
$i=2$	0 (6)	0 (3)	0 (5)	30 (15)	30
$i=3$	0 (2)	0 (1)	0 (2)	10 (5)	10
Total $J=j$	20	10	20	50	100

Table 8 Calculation of the weight (W) of the "error cells" and the weighted frequency (W·FE) of the expected number of model violations

Pair (i, j)	*FE*	δ_{j01}	δ_{j12}	δ_{i01}	δ_{j23}	δ_{i12}	δ_{i23}	*Weight (W)*	*W.FE*
0,3	20	1	1	0	1	0	0	1	20
1,0	4	0	0	1	0	0	0	2	8
1,1	2	1	0	1	0	0	0	1	2
2,0	6	0	0	1	0	1	0	5	30
2,1	3	1	0	1	0	1	0	3	9
2,2	5	1	1	1	0	1	0	1	5
3,0	2	0	0	1	0	1	1	8	16
3,1	1	1	0	1	0	1	1	3	3
3,2	2	1	1	1	0	1	1	2	4
Total	45								97

dichotomous variables. For each of these new response patterns we can calculate the number of item step pairs in which the "more difficult" item step receives the positive response and the "easier" item step receives the negative response. This number is called W (for weight as a measure of the seriousness of the violation, the factor with which to multiply the errors in the response to the original pair of items). Table 8 shows how this number is found for the original item pairs that contain a model error. It gives the error frequencies expected under stochastic independence (as FE—frequency of expected error). The product of W and FE ($W \cdot FE$) gives the weighted number of model violations expected under stochastic independence. A similar calculation yields the observed number of model violations, and a weighted coefficient of homogeneity can then be calculated using the weighted number of (observed and expected) errors ($H = 1 - E(\text{obs, weighted})/E(\text{exp, weighted})$. Molenaar (1991) has proven that this way of defining homogeneity is identical to a definition based on the ratio of the covariance between the two items over the maximal possible covariance, given the marginal frequency distribution of the items.

In addition to the concepts of model violation and homogeneity, most other aspects of the Mokken scaling procedure can also be easily extended to polytomous items. Coefficients of homogeneity for the whole scale and for individual items are calculated by either summing the number of weighted errors, observed and expected, or the covariances between the pairs of items and their maximal values. As extensions of the Z-statistic described for the dichotomous case, Z-statistics have also been developed to test whether the homogeneity coefficients in the population may still be 0. The search procedure that selects items to form a Mokken scale for polytomous items is identical to the procedure described for dichotomous items.

In the case of polytomous items the tests for Monotone Homogeneity and the tests for Double Monotonicity are carried out for each of the dichotomous item step variables separately. Molenaar's statistic CRIT now applies to the combination of outcomes of the model tests for the item step variables. Molenaar and Sijtsma (1988) have generalized Mokken's reliability coefficient for use with polytomous items, and Sijtsma (1998) has shown that a coefficient of item reliability, ρ_i, is identical to the coefficient of homogeneity for two independent replications of item i.

5 Examples of Mokken Scales Based on the 1990 World Values Study

Now we will turn to two examples that demonstrate the procedures described above, first for dichotomous data, and then for polytomous data. The data come from the World Values Study 1990, which was carried out in a large number of countries (Inglehart 1997). In each of these examples the first question to be addressed is whether the items can be ordered in a theoretically interesting way, i.e., a way that helps us understand the content of the latent trait we want to study. If so, then it is essential to use a measurement model that assumes a (cumulative) ordering, rather than models such as reliability or factor analysis that treat the items as clusters of unordered variables. A second question to be examined is whether the order of the items is constant, regardless of the sample (is sample-invariant). If the items are ordered in different ways for different subsets of respondents, then a strict model like the Rasch model is not appropriate.

5.1 *Religious Beliefs*

Let us start with questions about religious values, using the data from the United States, Canada, Ireland, and Sweden. The last two countries were selected as among the most (Ireland) and least (Sweden) religious countries in the European Union. For all the countries we will distinguish between those respondents who consider themselves as religious (value 1), and those who do not (value 0). The proportions of these respondents are given in Table 9. We do find a clear order in the percentage of respondents who hold a specific religious belief: "Belief in God" is mentioned most often (except among the nonreligious Swedes, who mention "Belief in a Soul" more often). For most samples "Belief in a Soul"

Table 9 Proportion of respondents in the religious or nonreligious subgroup of respondents (+Rel: religious; −Rel: not religious) who gave the positive ("religious") response to an item

		USA		*Canada*		*Ireland*		*Sweden*	
Belief in...	*Entire group*	*+Rel*	*−Rel*	*+Rel*	*−Rel*	*+Rel*	*−Rel*	*+Rel*	*−Rel*
Hell	0.56	0.81	0.46	0.53	0.23	0.66	0.43	0.32	0.03
Devil	0.56	0.80	0.46	0.55	0.23	0.67	0.49	0.40	0.05
Life after Death	0.72	0.86	0.50	0.79	0.45	0.90	0.69	0.71	0.24
Heaven	0.76	0.94	0.61	0.84	0.42	0.96	0.72	0.68	0.16
Sin	0.77	0.94	0.72	0.81	0.53	0.93	0.76	0.62	0.19
A Soul	0.83	0.96	0.71	0.92	0.61	0.94	0.73	0.82	0.40
God	0.87	0.99	0.79	0.98	0.64	1.00	0.90	0.91	0.27
N (group size)	10306	3905	648	1677	556	1269	491	544	1216

Reading example: 81% of the 3905 religious respondents in the American sample gave the positive response to the item "Belief in Hell" (i.e., they believe in Hell).

Table 10 Homogeneity analysis of seven religious belief items

	All respondents (N = 10,306, H = 0.76, RHO = 0.91)		*Discriminating respondents (N = 4828, H = 0.45, RHO = 0.66)*	
Belief in...	*Mean*	*ItemH*	*Mean*	*ItemH*
Hell	0.56	0.85	0.21	0.60
Devil	0.56	0.83	0.22	0.51
Life after Death	0.72	0.64	0.55[a]	0.13[a]
Heaven	0.76	0.76	0.64	0.51
Sin	0.77	0.69	0.66	0.36
A Soul	0.83	0.74	0.80	0.28
God	0.87	0.81	0.87	0.47

Note. $H = 0.76$, RHO = 0.91 ($N = 10{,}306$).
[a]Not in scale.

is more common than "Belief in Heaven," although for the Irish respondents the opposite is true. In each sample the least popular beliefs are "Belief in the Devil" and "Belief in Hell." There is, therefore, a theoretically interesting order of the items that helps us understand the content of the latent trait.

Can we use these questions to measure a single latent trait, religious belief, that holds for different (sub)populations? A Mokken scale analysis of the data for the entire group of 10,306 respondents shows an H coefficient of $H = 0.76$, which implies a very strong scale. The probability that this value would be found for a population in which the H-value is actually 0 is negligible ($Z = 245$). All item-coefficients of homogeneity are over 0.60, and all $Z(i)$ values are over 120. In Table 10 the items are ordered according to their mean value (i.e., the proportion of respondents who gave the "religious response") from low to high. The different popularity of the items suggests that a person who gives the positive (religious) response to any one item is also most likely to give the positive response to all the items listed below it.

The check of single monotonicity for the entire group, based on the rest score groups, reveals no violations (of .03 proportion points or more). Checks of double monotonicity show some significant violations in a few items (Belief in Heaven, Belief in Life after Death, Belief in Sin), but none of these reach CRIT-values that warrant deleting any of the items in the scale. Separate analyses for each of the subgroups also give high homogeneity values and no serious violations of single or Double Monotonicity. The reliability of the scale is 0.91.

The fact that different tests of the assumption of Double Monotonicity find no serious violations suggests that the scale will also conform to the requirements of a Rasch scale. There is, however, an important difference in the data used in a Rasch analysis from data used in a Mokken analysis: data from subjects whose responses do not discriminate between items (i.e., who give either the "religious" or the "nonreligious" response to all items) are treated differently—they are not used in conditional maximum likelihood item parameter estimation, and they receive an extrapolated scale value. In a Mokken analysis the user has the option to use or not to use these responses.

In this data set, subjects who give the same response to all items contribute heavily to the amount of correlation among the items. The mean inter-item correlation drops from

Table 11 Results of a Rasch scale analysis: item and person location parameters with their standard error, and two statistics for item fit

Belief in...	*Location (SE = .04)*	$\chi^2(df=5)$	*Residual*	*Person score*	*Location*	*SE*
				0	−3.83	2.01
God	−2.13	85	−3.08	1	−2.43	1.18
A Soul	−1.42	34	2.43	2	−1.33	0.97
Sin	−0.44	183	6.72	3	−0.46	0.92
Heaven	−0.36	256	−10.80	4	0.39	0.93
Life after Death	0.39	1183	17.51	5	1.32	1.00
Devil	1.87	349	−4.07	6	2.48	1.20
Hell	2.08	428	−7.65	7	3.94	1.74

0.53 to 0.31 if the respondents who give the same response to all seven items are dropped from the analysis. Still, Cronbach's α is a sizable 0.76. A second Mokken scale analysis, now performed on the 4828 respondents who discriminated in their responses, still gives acceptable results: The homogeneity of the scale drops to $H = 0.45$ (see Table 10). "Belief in Life after Death" is the only item that drops out of the scale, because it correlates negatively with "Belief in Hell" and "Belief in the Devil." Deleting this item leads to an acceptable final result.

Finally, a Rasch analysis, using the RUMM2010 program on the 4828 discriminating respondents, confirms our expectations that this scale forms an excellent Rasch scale, and that the item "Belief in Life after Death" has the worst fit (see Table 11). The index of person separation is 0.92; the mean fit statistic for the items is 0.151, and for the persons −0.283. [Note: the user manuals (Sheridan et al. 2000, 2001) unfortunately give us little help in interpreting these values.] Because the respondents are measured with only seven items, the standard error of their location parameter is so high that their apparent locations can serve only as estimates of their rank order along the latent dimension.

Why does the item "Belief in Life after Death" not pass the scrutiny of the tests? A researcher who wants to measure the degree of religious belief in respondents in different countries, and who has selected these items, will need to find an answer to this question. The difference in the order with which this belief is held by different subgroups of respondents may provide a clue as to where to look for an answer.

5.2 *Confidence in National Institutions*

To illustrate the analysis of polytomous data, I have selected nine questions from the World Values Study 1990 about the amount of confidence people have in nine national institutions: trade unions, parliament, the civil service, the press, major companies, the legal system, the police, the armed forces, and the church. The possible responses were "a great deal" (3); "quite a lot" (2); "not very much" (1); and "none at all" (0). It is likely that the amount of confidence will differ systematically across institutions and across respondents, which means that combining their responses to different items would allow us to identify their position along a latent trait that we might call "extent of confidence in national institutions."

As the relative amount of confidence in various national institutions may differ across countries, we will use respondents from four countries: two in North America (the United States and Mexico) and two in the European Union (Italy and Austria). Respondents differ in their political ideology, as indicated by their self-placement on a left–right scale (leftist:

Table 12 Mean values of 10 groups, based on country and left–right self placement (1–4 = left, 5 = center, 6–10 = right) in their answers about degree of confidence in 9 national institutions

	Entire group	*USA*			*Mexico*			*Italy*		*Austria*	
		L	*C*	*R*	*L*	*C*	*R*	*L*	*R*	*L*	*R*
Unions	1.24	1.42	1.29	1.27	1.15	1.08	1.30	1.26	0.98	1.29	1.21
Parliament	1.29	1.41	1.40	1.49	1.10	1.06	1.26	1.04	1.15	1.44	1.38
Civil Services	1.31	1.62	1.64	1.66	0.96	0.99	1.15	0.88	1.08	1.46	1.41
Press	1.36	1.49	1.43	1.45	1.37	1.32	1.46	1.23	1.25	0.92	1.01
Companies	1.42	1.43	1.52	1.60	1.21	1.22	1.38	1.22	1.64	1.38	1.40
Legal system	1.49	1.51	1.52	1.60	1.31	1.37	1.63	1.14	1.27	1.64	1.67
Police	1.59	1.81	1.91	2.01	0.92	0.97	1.25	1.50	1.87	1.76	1.80
Army	1.64	1.83	1.98	2.08	1.35	1.39	1.62	1.09	1.66	1.07	1.17
Church	1.97	1.83	2.12	2.25	1.81	2.00	2.21	1.19	1.98	1.40	1.60
n	11680	799	1443	2311	654	1047	2433	1110	751	140	992

Note. Responses could be 3 = "a great deal"; 2 = "quite a lot"; 1 = "not very much"; and 0 = "none at all."

score 1–4; rightist: score 6–10). As there were many American and Mexican respondents who chose the category 5, I placed these respondents in a separate "middle" category. For each of the resulting 10 groups (2 countries with 2 ideological categories and 2 countries with 3 ideological categories) the mean response for each of the 9 national institutions is given in Table 12.

Institutions clearly differ in how much confidence they inspire. Some, notably the trade unions and the parliament, are perceived with little confidence, whereas others, especially the Church, enjoy a lot. It is appropriate, then, to analyze these data with an IRT measurement model that takes these differences into account.

Is the order in which the institutions are perceived as inspiring confidence the same across all subgroups of respondents? There are systematic differences between the countries, as Table 12 shows. Within each country, right-wing respondents have more confidence on the whole in most of the institutions than left-wing respondents. Conversely, for all countries but Mexico, left-wingers have more confidence in trade unions than do right-wingers. This is the exception, however, because within each country the order of most of the other institutions is the same for left- and right-wing respondents. The differences in the ordering of these institutions in the different subgroups look serious enough that the assumption of Invariant Item Ordering is likely to be violated, and so neither the Double Monotonicity model nor the Rasch model will fit very well.

A Mokken scale analysis on the entire data set confirms this. Eight of the nine institutions form an acceptable cumulative scale (with a weighted H-value of 0.40), but the item-homogeneity of Confidence in the Church falls below the lower boundary of 0.30. To test the model of Monotone Homogeneity, subgroups of at least 100 respondents were formed, and violations of at least .03 proportion points were inspected. Less than 1% of the (more than 500) appropriate comparisons lead to significant violations. The overall evaluation of all items in the scale is good (the highest CRIT value, 42, is well within the acceptable zone). The model of Double Monotonicity shows serious violations in each of the four tests, except in the comparison of the item step order for subgroups of respondents on the basis of their left–right self-placement.

Mokken scale analyses were then performed on the data from each country separately. For all countries an acceptable cumulative scale was found, but with different subsets of

Table 13 Mokken scaling (homogeneity analysis) results on amount of confidence in national institutions in the United States, Mexico, Italy, and Austria

Label	*Mean*	*ItemH*
USA: $H = 0.38$, RHO $= 0.75$ *(*$n = 4553$*)*		
Confidence in...		
Parliament	1.45	0.43
Companies	1.54	0.36
Legal system	1.56	0.33
Civil Service	1.65	0.43
Police	1.94	0.43
Armed forces	2.01	0.31
Nonscaling items because of insufficient homogeneity:		
Unions	1.30	0.26
Press	1.45	0.27
Church	2.14	0.25
Mexico: $H = 0.50$, RHO $= 0.87$ ($n = 4134$)		
Confidence in ...		
Civil Service	1.08	0.57
Police	1.13	0.54
Parliament	1.18	0.56
Unions	1.22	0.50
Companies	1.32	0.47
Press	1.41	0.51
Legal system	1.51	0.47
Armed forces	1.52	0.42
Church	2.09	0.27
Italy: $H = 0.40$, RHO $= 0.77$ ($n = 1861$)		
Confidence in...		
Civil Service	0.96	0.41
Parliament	1.09	0.40
Legal system	1.19	0.38
Armed forces	1.32	0.40
Church	1.51	0.36
Police	1.65	0.47
Nonscaling items because of insufficient homogeneity:		
Unions	1.14	0.19
Press	1.24	0.23
Companies	1.39	0.28
Austria: $H = 0.41$, RHO $= 0.77$ ($n = 1132$)		
Confidence in...		
Armed forces	1.16	0.57
Parliament	1.38	0.54
Companies	1.40	0.56
Civil Service	1.41	0.50
Legal system	1.67	0.47
Police	1.79	0.51
Press	1.00	0.25
Unions	1.22	0.29
Church	1.57	0.28

items, and with the items in different order of "popularity" (see Table 13). For each of the four scales, both the test for Monotone Homogeneity and all tests for Double Monotonicity showed the amount of model violation to be within the acceptable range.

6 Discussion

Measurement models like reliability or factor analysis assume that all items are equally "popular" (i.e., have the same frequency distribution). Whenever this assumption is violated—as when the items form a cumulative scale—an artifact can creep in whereby items do not seem to be homogeneous enough to measure a single latent variable. In these situations a major advantage of IRT models, including Mokken scale analysis, over CTT models is that in introducing model parameters for items, they explicitly take into account that the items differ in popularity.

A second advantage of Mokken scale analysis over CTT models lies in the detailed emphasis on model fit. All H_{ij} coefficients (and therefore all pairwise correlations) must be positive, and each item must be sufficiently homogeneous with the others. These requirements lead to measurement instruments that conform to higher standards of reliability and homogeneity than instruments that have been inspected only in a standard reliability analysis.

A third, more practical, advantage of the Mokken scaling procedure over most other measurement procedures is its "bottom-up" hierarchical clustering search procedure that identifies a maximal subset of homogeneous items. Especially in exploratory research aimed at developing new measurement instruments, this procedure helps the researcher to detect new candidates for latent variables, even when only a limited number of items are available.

Finally, Mokken scale analysis is an IRT model that can successfully be used on small numbers of items. Molenaar (1997a) showed that when the number of items is relatively small, the results of a Mokken scale analysis and the more stringent Rasch scale analysis often lead to essentially the same results.

Software for nonparametric and parametric IRT models, including Mokken scale analysis, is available as Windows-oriented, well-documented, and user-friendly stand-alone software.[6] Appendix 1 (available on the *Political Analysis* Web site) gives some additional information that may be useful for readers who want to perform a Mokken scale analysis on their own data.

The Mokken models have led to the developments of further nonparametric IRT models that differ from Guttman's original cumulative model in the specification of their IRF. Two new approaches, based on Mokken's scaling models and procedure, have been particularly fruitful: the construction of models and procedures for nonparametric unfolding analysis (Post and Snijders 1993; Van Schuur 1993, 1997, 1998; Van Schuur and Kiers 1994), and for the nonparametric circumplex (Mokken et al. 2001).

Mokken scale analysis has been applied for more than 30 years. Sijtsma and Molenaar (2002) give an overview of applications, mainly in testing and survey research. For applications in political science (many not involving surveys), consult Cingranelli and Richards

[6] Software for Mokken scale analysis is available as the program *MSP5 for Windows* from ProGamma; e-mail: gamma.post@gamma.rug.nl; Web site: http//www.gamma.rug.nl. This program supercedes previous versions such as *SCAMMO* or *Mokken Scale*. Other algorithms, at least in limited form, are available as SAS macros and STATA ADO files, or as stand-alone programs (cf. Kingma and ten Vergert 1985). Software for parametric IRT models is also available (see Embretson and Reise 2000, ch. 13, for an overview). For the Rasch analyses discussed in this paper I used the *RUMM2010* program (Sheridan et al. 2000, 2001), which performs a Rasch analysis on dichotomous data and the Partial Credit Model and the Rating Scale Model analysis on polytomous data. This program is available from www.rummlab.com.au; e-mail: rummlab@arach.net.au.

(1999), Davenport (1995), Jacoby (1994, 1995), Mokken (1971), Richards, et al. (2001), Scarritt (1996), Schneider et al. (1997), Stokman (1977), Van Schuur and Vis (2000), and Zin et al. (1992).

References

Andrich, D. 1988. *Rasch Models for Measurement.* Newbury Park, CA: Sage.

Andrich, D., and G. A. Douglas. 1977. "Reliability: Distinctions Between Item Consistency and Subject Separation Within the Simple Logistic Model." Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Bart, W. M., and D. J. Krus. 1973. "An Ordering Theoretic Method to Determine Hierarchies Among Items." *Educational and Psychological Measurement* 33:291–300.

Birnbaum, A. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In *Statistical Theories of Mental Test Scores*, eds. F. M. Lord and R. Novick. Reading, MA: Addison-Wesley.

Carroll, J. B. 1945. "The Effect of Difficulty and Chance Success on Correlations Between Items or Between Tests." *Psychometrika* 10:1–19.

Cingranelli, D. L., and D. L. Richards. 1999. "Measuring the Level, Pattern and Sequence of Government Respect for Physical Integrity Rights." *International Studies Quarterly* 43:407–417.

Coombs, C. H., and J. C. Lingoes. 1978. "Stochastic Cumulative Scales." In *Theory Construction and Data Analysis in the Behavioral Sciences*, ed. S. Shye. San Francisco: Jossey-Bass, pp. 280–298.

Davenport, C. 1995. "Multidimensional Threat Perception and State Repression: An Inquiry Into Why States Apply Negative Sanctions." *American Journal of Political Science* 39:683–713.

Dayton, C. M., and G. B. MacReady. 1980. "A Scaling Model with Response Errors and Intrinsically Unscalable Respondents." *Psychometrika* 45:343–356.

Embretson, S., and S. P. Reise. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Ferguson, G. A. 1941. "The Factorial Interpretation of Test Difficulty." *Psychometrika* 6:323–330.

Ganter, B., and R. Wille. 1999. *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer-Verlag.

Guttman, L. 1950. "The Basis for Scalogram Analysis." In *Measurement and Prediction. Studies in Social Psychology in World War II,* vol. 4, eds. S. A. Stouffer et al. Princeton, NJ: Princeton University Press, pp. 60–90.

Inglehart, R. 1997. *Modernization and Postmodernization: Cultural, Economic and Political Change in 43 Countries.* Princeton, NJ: Princeton University Press.

Jacoby, W. G. 1994. "Public Attitudes Towards Government Spending." *American Journal of Political Science* 38:336–361.

Jacoby, W. G. 1995. "The Structure of Ideological Thinking in the American Electorate." *American Journal of Political Science* 39:314–335.

Kingma, J., and E. ten Vergert. 1985. "A Nonparametric Scale Analysis of the Development of Conservation." *Applied Psychological Measurement* 9:375–387.

Loevinger, J. 1948. "The Technique of Homogeneous Tests Compared with Some Aspects of 'Scale Analysis' and Factor Analysis." *Psychological Bulletin* 45:507–530.

Meijer, R. R. 1994. *Nonparametric Person Fit Analysis*. Unpublished doctoral dissertation. Amsterdam: Free University.

Meijer, R. R., and K. Sijtsma. 2001. "Methodology Review: Evaluating Person Fit." *Applied Psychological Measurement* 25:107–135.

Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis with Applications in Political Research.* New York: De Gruyter.

Mokken, R. J. 1997. "Nonparametric Models for Dichotomous Responses." In *Handbook of Modern Item Response Theory*, eds. W. J. van der Linden and R. K. Hambleton. New York: Springer-Verlag, 351–367.

Mokken, R. J., and C. Lewis. 1982. "A Nonparametric Approach to the Analysis of Dichotomous Item Responses." *Applied Psychological Measurement* 6:417–430.

Mokken, R. J., W. H., van Schuur, and A. J. Leeferink. 2001. "The Circles of Our Minds. A Nonparametric IRT Model for the Circumplex." In *Essays on item response theory*, eds. A. Boomsma, M. A. J. van Duijn, and T. A. B. Snijders. New York: Springer-Verlag, pp. 339–356.

Molenaar, I. W. 1973. "Simple Approximations to the Poisson, Binomial and Hypergeometrical Distributions." *Biometrics* 29:403–407.

Molenaar, I. W. 1991. "A Weighted Loevinger H Coefficient Extending Mokken Scaling to Multicategory Items." *Kwantitatieve Methoden* 12:97–117.

Molenaar, I. W. 1997a. "Lenient or Strict Application of IRT with an Eye on Practical Consequences." In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, eds. J. Rost and R. Langeheine. Münster: Waxmann, pp. 38–49.

Molenaar, I. W. 1997b. "Nonparametric Models for Polytomous Responses." In *Handbook of Modern Item Response Theory*, eds. W. J. van der Linden and R. K. Hambleton. New York: Springer-Verlag, pp. 367–380.

Molenaar, I. W. and K. Sijtsma. 1988. "Mokken's Approach to Reliability Estimation Extended to Multicategory Items." *Kwantitatieve Methoden* 9:115–126.

Molenaar, I. W., and K. Sijtsma. 2000. *MSP5 for Windows. A Program for Mokken Scale Analysis for Polytomous Items.* Groningen: ProGamma.

Niemöller, B., and W. H. van Schuur. 1983. "Stochastic Models for Unidimensional Scaling: Mokken and Rasch." In *Data Analysis and the Social Sciences*, eds. D. McKay, N. Schofield, and P. Whiteley. London: Francis Pinter, pp. 120–170.

Post, W. J., and T. A. B. Snijders. 1993. "Nonparametric Unfolding Models for Dichotomous Data." *Methodika* 7:130-156.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Nielsen and Lydiche.

Richards, D. L., R. D., Gelleny, and D. H. Sacko. 2001. "Money with a Mean Streak? Foreign Economic Penetration and Government Respect for Human Rights in Developing Countries." *International Studies Quarterly* 45:219–231.

Rosenbaum, P. R. 1984. "Testing the Conditional Independence and Monotonicity Assumptions of Item Response Theory." *Psychometrika* 49:425–435.

Rosenbaum, P. R. 1987. "Comparing Item Characteristic Curves." *Psychometrika* 52:217–233.

Samejima, F. 1969. "Estimation of Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph* 17:1–100.

Scarritt, J. R. 1996. "Measuring Political Change: The Quantity and Effectiveness of Electoral and Party Participation in the Zambian One-Party State, 1973–1991." *British Journal of Political Science* 26:283–297.

Schneider, S. K., W. G. Jacoby, and J. D. Coggburn. 1997. "The Structure of Bureaucratic Decision Making in the American States." *Public Administration Review* 57:240–249.

Schriever, B. F. 1985. *Order Dependence.* Unpublished doctoral dissertation. Amsterdam: Free University Press.

Sheridan, B., D. Andrich, and G. Luo. 2000. *RUMM2010 Manual, Part 2: Extending RUMM2010*. Duncraig, Western Australia: RUMM Laboratory.

Sheridan, B., D. Andrich, and G. Luo. 2001. *RUMM2010 Manual, Part 1: Getting Started.* Duncraig, Western Australia: RUMM Laboratory.

Shye, S. 1985. *Multiple Scaling*. Amsterdam: North Holland.

Sijtsma, K. 1998. "Beyond Mokken Scale Analysis." In *In Search of Structure. Essays in Social Science and Methodology*, eds. M. Fennema, C. van der Eijk, and H. Schijf. Amsterdam: Het Spinhuis, pp. 29–44.

Sijtsma, K., and I. W. Molenaar. 1987. "Reliability of Test Scores in Nonparametric Item Response Theory." *Psychometrika* 52:79–97.

Sijtsma, K., and I. W. Molenaar. 2002. *Introduction to Nonparametric Item Response Theory.* Vol. 5 of *Measurement Methods for the Social Sciences*. Thousand Oaks, CA: Sage.

Stokman, F. N. 1977. *Roll Calls and Sponsorship: A Methodological Analysis of Third World Group Formation in the United Nations.* Leiden: Sijthoff.

Van Schuur, W. H. 1993. "Nonparametric Unfolding Models for Multicategory Data." *Political Analysis* 4:41–74.

Van Schuur, W. H. 1997. "Nonparametric IRT Models for Dominance and Proximity Data." In *Objective Measurement: Theory into Practice*, vol. 4, eds. M. Wilson, G. Engelhard Jr., and K. Draney. Greenwich, London: Ablex, pp. 313–331.

Van Schuur, W. H. 1998. "From Mokken to Mudfold and Back." In *In Search of Structure. Essays in Social Science and Methodology*, eds. M. Fennema, C. van der Eijk, and H. Schijf. Amsterdam: Het Spinhuis, pp. 45–62.

Van Schuur, W. H., and H. A. L. Kiers. 1994. "Why Factor Analysis is Often the Wrong Model for Analyzing Bipolar Concepts, and What Model to Use Instead." *Applied Psychological Measurement* 18:97–110.

Van Schuur, W. H., and J. C. P. M. Vis. 2002. "What Dutch Parliamentary Journalists Know About Politics." *Acta Politica* 35:196–227.

Zinn, F. D., D. A. Henderson, J. D. Nystuen,and W. D. Drake. 1992. "A Stochastic Cumulative Scaling Method Applied to Measuring Wealth in Indonesian Villages." *Environment and Planning A* 24:1155–1166.