Selection of Unidimensional Scales From a Multidimensional Item Bank in the Polytomous Mokken IRT Model

Bas T. Hemker and Klaas Sijtsma, Utrecht University

Ivo W. Molenaar, University of Groningen

An automated item selection procedure for selecting unidimensional scales of polytomous items from multidimensional datasets is developed for use in the context of the Mokken item response theory model of monotone homogeneity (Mokken & Lewis, 1982). The selection procedure is directly based on the selection procedure proposed by Mokken (1971, p. 187) and relies heavily on the scalability coefficient H (Loevinger, 1948; Molenaar, 1991). New theoretical results relating the latent model structure to H are provided. The item selection procedure requires selection of a lower bound for H. A simulation study determined ranges of H for which the unidimensional item sets were retrieved from multidimensional datasets. If multidimensionality is suspected in an empirical dataset, well-chosen lower bound values can be used effectively to detect the unidimensional scales. *Index terms: item response theory, Mokken model, multidimensional item banks, nonparametric item response models, scalability coefficient H, test construction, unidimensional scales.*

Multidimensionality manifests itself in different ways in test construction research. For example, for test batteries that address complex multidimensional constructs, such as intelligence, subsets of items can be constructed for each relevant aspect of the construct. Data obtained from these test batteries are multidimensional by construction. However, subsets of items for subtests in such a test battery are intended to be unidimensional.

When constructing a test for a unidimensional construct, multidimensionality in a dataset may be introduced in several ways. For example, due to an unfortunate phrasing a few items may measure other traits than the majority of the items. Another example is that the test constructor intended the test to include a number of different substantive areas, but was not aware of the resultant multidimensionality that was introduced.

Many methods have been used to identify unidimensional item sets from larger item banks (e.g., exploratory factor analysis and cluster analysis). Another approach to analyzing multidimensionality is to use multidimensional item response theory (IRT) models (e.g., Batley & Boss, 1993; Reckase & McKinley, 1991). Rather than selecting items into subsets, that approach describes the latent structure by means of item parameters and a person parameter for each latent trait.

An iterative method for selecting unidimensional scales from a multidimensional item bank is investigated here. This method is subsumed under the Mokken (1971; Mokken & Lewis, 1982) nonparametric IRT approach to scaling. The item selection procedure selects from an item bank items that approximately satisfy the requirements of the Mokken IRT model of monotone homogeneity. After one set of items that forms a scale has been selected, the item selection procedure attempts to find the next scale from the subset of unselected items, and so on, until there are no more items left that can form a monotonely homogeneous scale.

Compared with factor analysis and related methods, a definite advantage of this item selection procedure is that items are selected on the basis of their fit to a particular IRT model. Compared with multidimensional

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 4, December 1995, pp. 337–352

0146-6216/95/040337-16$2.05

IRT models, the advantage may be that the interpretation of test performance on the basis of scores obtained on unidimensional tests is relatively straightforward.

The Mokken Approach to Scaling

Assumptions and Definitions

In IRT models, the probability of a particular response to an item is a function of characteristics of the person and the item. For dichotomous items, this relationship is expressed by the item response function (IRF). Nonparametric IRT models (e.g., Junker, 1993; Mokken, 1971; Mokken & Lewis, 1982; Rosenbaum, 1987; Stout, 1990) do not assume a parametric definition of the IRFs or of the latent trait (θ) distribution across people. These models thus are based on weaker assumptions than most parametric IRT models. As a result, nonparametric models yield only ordinal measurement, whereas parametric models provide interval measurement. However, nonparametric models often fit empirical data better than parametric models (Meijer, Sijtsma, & Smid, 1990; Mokken & Lewis, 1982; Molenaar, in press). Meijer et al. (1990) and de Gruijter (1993) provided a comparison of some nonparametric and parametric IRT models.

This study used the nonparametric Mokken IRT model for polytomous items (Molenaar, 1982, 1986, in press). Let X_i denote the random variable for the score on item i ($i = 1, ..., k$) with $m + 1$ ordered answer categories. A given item score is denoted $X_i = x_i$, where $x_i = 0, ..., m$. An item step is the imaginary threshold between two adjacent ordered answer categories. Each item with $m + 1$ categories is assumed to be based on m hypothetical dichotomous item steps indexed by $s = 1, ..., m$. Going from a lower to a higher response category yields 1 point; otherwise a score of 0 is assigned. Within one item, an item step cannot be passed if an earlier item step has not been passed. A 0 step score thus implies 0 step scores on the next steps of the same item, and a step score of 1 implies scores of 1 on preceding steps.

Item step scores are denoted by Y_{is}. The relationship between X_i and Y_{is} is

$$X_i = \sum_{s=1}^{m} Y_{is}. \tag{1}$$

The probability of successfully taking step s of item i is written as $P(Y_{is} = 1|\theta)$, which is equivalent to $P(X_i \geq s|\theta)$; the notation $\pi_{is}(\theta)$ also is used. This conditional probability is the item step response function (ISRF). For dichotomous items, $m = 1$ and the ISRF is identical to the IRF.

The Mokken model of monotone homogeneity is defined by three assumptions: (1) unidimensionality, which means that all items measure the same trait; (2) local stochastic independence, which implies that for fixed θ the covariance between item scores X_i and X_j ($i = 1, ..., k$; $j = 1, ..., k$; $i \neq j$) is 0; and (3) monotonicity in θ, which means that for every item i the following proposition is true: Let person A and person B have latent trait values θ_A and θ_B with $\theta_A < \theta_B$, then

$$P(Y_{is} = 1|\theta_A) \leq P(Y_{is} = 1|\theta_B) \tag{2}$$

for all pairs of different latent trait values, which means that the ISRFs are nondecreasing. A similar inequality also holds for the total score X on k polytomous items with m answer categories each. Note that $P(Y_{is} = 1|\theta) = E(Y_{is}|\theta)$. Summation across km item steps would yield:

$$\sum_{i=1}^{k}\sum_{s=1}^{m} E(Y_{is}|\theta_A) \leq \sum_{i=1}^{k}\sum_{s=1}^{m} E(Y_{is}|\theta_B) \tag{3}$$

or, equivalently,

$$\mathrm{E}\left(\sum_{i=1}^{k}\sum_{s=1}^{m} Y_{is}|\theta_{\mathrm{A}}\right) \leq \mathrm{E}\left(\sum_{i=1}^{k}\sum_{s=1}^{m} Y_{is}|\theta_{\mathrm{B}}\right), \tag{4}$$

and

$$\mathrm{E}(X|\theta_{\mathrm{A}}) \leq \mathrm{E}(X|\theta_{\mathrm{B}}). \tag{5}$$

Note that this result also holds if m varies across items.

Relationship With Other Models

Analogous to Mokken's (1971) approach to dichotomous items, a more restrictive model than the model of monotone homogeneity can be formulated for polytomous items. This more restrictive model can be designated the model of double monotonicity for polytomous items (e.g., Molenaar, in press). This model is based on the model of monotone homogeneity and, in addition, requires that the ISRFs of different items do not intersect (Mokken & Lewis, 1982; Molenaar, in press). Thus, the model of double monotonicity is nested in the model of monotone homogeneity (Molenaar, 1986, in press). The model of monotone homogeneity suffices for many applications.

The model of monotone homogeneity for polytomous items is a nonparametric version of the graded response model (Masters, 1982; Samejima, 1969). This is obvious from the following line of reasoning. Let α_i denote the item discrimination parameter, and let λ_{is} denote the category boundary parameter of step s of item i. Then, according to the graded response model,

$$P(X_i \geq s|\theta) = \frac{\exp[\alpha_i(\theta - \lambda_{is})]}{1 + \exp[\alpha_i(\theta - \lambda_{is})]}. \tag{6}$$

The category boundary parameter λ_{is} can be written as the sum of the difficulty of item i, δ_i, and the difficulty of step s of item i, τ_{is}, with

$$\sum_{s=1}^{m} \tau_{is} = 0 \tag{7}$$

and

$$\delta_i = \left(\sum_{s=1}^{m} \lambda_{is}\right) \Big/ m. \tag{8}$$

The probability of the item score $X_i = x_i$, $P(X_i = x_i|\theta)$, can be obtained from these ISRFs. If $X_i = s$ then

$$P(X_i = s|\theta) = P(Y_{is} = 1|\theta) - P(Y_{i,s+1} = 1|\theta), \tag{9}$$

with $P(Y_{i,m+1} = 1|\theta) = 0$ and $P(Y_{i0} = 1|\theta) = 1$. Note that the formulation of item steps in the nonparametric model of monotone homogeneity for polytomous items implies the same relation between probabilities of item scores and probabilities of item step scores.

Observable Consequences

If the model of monotone homogeneity holds for dichotomous items, the covariance σ_{ij} between items i and j is non-negative (Mokken, 1971, pp. 120, 131; see also Ellis & van den Wollenberg, 1993; Mokken & Lewis, 1982). This result also holds for polytomous items.

Theorem. If monotone homogeneity holds, then $\sigma(X_i, X_j) \geq 0$; $X_i, X_j = 0, 1, \ldots, m$; $i \neq j$.

Proof. Assume a doubly stochastic process. First, θs are randomly sampled from the population distri-

bution. Second, given this sample, for each θ, replications of X_i and X_j are randomly sampled from the propensity distributions (Lord & Novick, 1968, p. 30) of the sampled persons. The covariance between X_i and X_j is

$$\sigma(X_i, X_j) = \sigma_\theta\left[\mathrm{E}(X_i|\theta), \mathrm{E}(X_j|\theta)\right] + \mathrm{E}_\theta\left[\sigma(X_i, X_j|\theta)\right]. \tag{10}$$

Because of local stochastic independence, the second term on the right equals 0. Next, note that because

$$X_i = \sum_{s=1}^{m} Y_{is}, \tag{11}$$

and $\mathrm{E}(Y_{is}|\theta) = \pi_{is}(\theta)$, if the item steps of item i are indexed by s and those of item j by t, the first term on the right can be written as

$$\sigma_\theta\left[\mathrm{E}\left(\sum_{s=1}^{m} Y_{is}|\theta\right), \mathrm{E}\left(\sum_{t=1}^{m} Y_{jt}|\theta\right)\right] = \sum_{s=1}^{m}\sum_{t=1}^{m} \sigma_\theta\left[\pi_{is}(\theta), \pi_{jt}(\theta)\right]. \tag{12}$$

Thus, from Equations 10 and 12 it follows that

$$\sigma(X_i, X_j) = \sum_{s=1}^{m}\sum_{t=1}^{m} \sigma_\theta\left[\pi_{is}(\theta), \pi_{jt}(\theta)\right]. \tag{13}$$

Next, note that by assumption, $\pi_{is}(\theta)$ and $\pi_{jt}(\theta)$ are monotonically nondecreasing functions of θ. Thus, these functions are similarly ordered functions (Mokken, 1971, p. 119). Mokken (p. 120, corollary 1.1.1) proved that the values of similarly ordered functions have a nonnegative covariance. Therefore, $\sigma_\theta[\pi_{is}(\theta), \pi_{jt}(\theta)] \geq 0$; thus, $\sigma(X_i, X_j) \geq 0$ follows.

Scalability

In the item selection procedure based on the model of monotone homogeneity, the scalability coefficient H (Loevinger, 1948; Molenaar, 1991) plays an important role. Given the number of items and the model of monotone homogeneity, the larger the value of H, the more accurately persons can be ordered. The property of non-negative values of σ_{ij} is fundamental for H in the context of the monotone homogeneity model.

From the covariance σ_{ij} and the maximum covariance $\sigma_{ij(\max)}$ given the marginals of the bivariate cross-tabulation of the scores on items i and j, H for these items is defined as

$$H_{ij} = \frac{\sigma_{ij}}{\sigma_{ij(\max)}} \tag{14}$$

[for dichotomous items refer to Mokken (1971) and Mokken & Lewis (1982); for polytomous items refer to Molenaar (1991)].

H_{ij} also can be defined as a decreasing function of the proportion of weighted Guttman (1950) errors, given that item step difficulties are fixed (Hemker & Sijtsma, 1993; Molenaar, 1991).

The extension of H_{ij} to the scalability coefficient H for k items is

$$H = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \sigma_{ij}}{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \sigma_{ij(\max)}}. \tag{15}$$

The extension to coefficient H_i, which indicates whether item i is scalable in accordance with the monotone

homogeneity model given the other items used, is

$$H_i = \frac{\sum_{j \neq i} \sigma_{ij}}{\sum_{j \neq i} \sigma_{ij(\max)}}. \tag{16}$$

It can be shown that $\min(H_{ij}) \leq \min(H_i) \leq H \leq \max(H_i) \leq \max(H_{ij})$. The proof of these inequalities is analogous to the proof for dichotomous items (Mokken, 1971, p. 152).

The null hypothesis that $H_{ij} = 0$, $H_i = 0$, or $H = 0$ against the alternative that H_{ij}, H_i, or H is positive can be tested for dichotomous items (Mokken, 1971, pp. 157–169) and polytomous items (e.g., Molenaar, in press; Molenaar, Debets, Sijtsma, & Hemker, 1994). In both cases, a test statistic z is used that is asymptotically standard normal. Because this test statistic was not used in this study, the technical details are not discussed here.

The maximum value of H is 1, which is obtained if all item pairs have maximum covariances. If the monotone homogeneity model holds, the minimum value for σ_{ij} is 0 and, consequently, the minimum value for H_{ij}, H_i, and H is also 0. Thus, a necessary condition for the monotone homogeneity model is $0 \leq H_{ij} \leq 1$ for all item pairs. Therefore, H_i and H also range from 0 to 1, given that the monotone homogeneity model holds. Note that if $\sigma_{ij} < 0$ then $H_{ij} < 0$, which contradicts the model.

Given the theorem, and the definition of H for polytomous items, the next corollary is obtained easily.

Corollary. For a set of k polytomous items satisfying monotone homogeneity $0 \leq H \leq 1$ with equality to 0 if and only if at least for $k - 1$ items the ISRFs are constant functions of θ.

Proof of necessity. Let $E[\pi_{is}(\theta)] = \pi_{is}$, $E[\pi_{jt}(\theta)] = \pi_{jt}$, and $i \neq j$, then

$$\sigma_\theta\left[\pi_{is}(\theta), \pi_{jt}(\theta)\right] = E_\theta\left\{\left[\pi_{is}(\theta) - \pi_{is}\right]\left[\pi_{jt}(\theta) - \pi_{jt}\right]\right\}. \tag{17}$$

Equation 17 shows that if all items have constant ISRFs [i.e., if $\pi_{is}(\theta) = \pi_{is}$ and if $\pi_{jt}(\theta) = \pi_{jt}$ for all θ], then all covariances between ISRFs are 0. This is also true if one item has one or more monotonically nondecreasing ISRFs, because Equation 17 pertains to ISRFs of different items. Thus, from Equation 13 it then follows that $\sigma_{ij} = 0$ for all item pairs ($i \neq j$). Next, from Equation 15 it follows that, given the monotone homogeneity model, $H = 0$ if at least $k - 1$ items have constant ISRFs.

Proof of sufficiency. $H = 0$ implies that the sum of all covariances in its numerator (see Equation 15) equals 0. Applying the theorem, this sum thus contains no negative covariances. Therefore, $H = 0$ only if all covariances equal 0, which means that $\sigma_{ij} = 0$ for all item pairs (i, j). The theorem shows that this means that all $\sigma_\theta[\pi_{is}(\theta), \pi_{jt}(\theta)] = 0$ for $i \neq j$. Given the model of monotone homogeneity, the ISRFs are either monotonically nondecreasing or constant. If at least one of two ISRFs is a constant function of θ, the covariance between these ISRFs is 0. Otherwise, given the monotone homogeneity model, this covariance is positive due to similarly orderedness (Mokken, 1971, p. 120, corollary 1.1.1). Thus, given the monotone homogeneity model and $H = 0$, at most one item can have one or more monotonically increasing ISRFs.

Because a positive H is not a sufficient condition for the monotone homogeneity model, and because low positive H values do not lead to useful scales allowing an accurate ordering of persons, Mokken (1971, p. 184) suggested the lower bound of $H = .30$ for practical use with dichotomous items. Scales with smaller H do not yield a satisfactory discrimination among persons. For the interpretation of the other values of H, Mokken (1971, p. 185) suggested the following guidelines:

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

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

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

The stronger the scale, the more accurately persons can be ordered. Hemker & Sijtsma (1993) concluded that these general rules also can be used for polytomous items because the number of answer categories has no substantial effect on H.

The use of these general rules thus prevents outcomes of item selection that are in agreement with the monotone homogeneity model but useless for practical purposes. An excellent example was given by Wood (1978) who found that data generated by coin flipping were in agreement with the Rasch model, which is a special case of the monotone homogeneity model for dichotomous items. Obviously, coin flipping represents an extreme case of both models in which all ISRFs are constant functions of θ and the total score has zero reliability. A lower bound for H forces the selection of items with at least moderate discriminations or reliability (Meijer, Sijtsma, & Molenaar, 1995). Note that this lower bound and the general rules are part of the scaling procedure, not of the model.

For dichotomous items, coefficient H is sensitive to the interplay of three other factors (Mokken, Lewis, & Sijtsma, 1986): the population variance, the slopes of the ISRFs (discriminations), and the spread of the item locations (difficulties). Holding constant two of these factors, H is an increasing function in the third factor. The same reasoning holds for polytomous items.

When the population variance is 0, there is no relevant information available regarding nontrivial monotone homogeneity of the item set. As a result, $H = 0$. As the population variance increases, holding constant the other two factors, more information is available about the ISRFs and, therefore, if the monotone homogeneity model holds, H will be larger.

In the boundary case of the monotone homogeneity model with smallest possible slopes (i.e., slopes equal to 0), all mk ISRFs are constant functions of θ, which results in $H = 0$ (see corollary). Increasing the slopes, while holding the other two factors constant, increases H. Maximum positive slopes yield ISRFs in agreement with the Guttman (1944) model; therefore, $H = 1$.

If the distance between all ISRFs is 0, the H value is a function of the population variance and the slopes. However, for positive, fixed slopes and given nonzero population variance, an increase of the distances between the ISRFs yields an increase of H. This is a result of the increase of the θ range over which the items have the potential of providing relevant information regarding monotone homogeneity.

Item Selection Procedure

Automated Item Selection

The bottom-up item selection procedure selects items from the initial set into scales that satisfy the requirements of the monotone homogeneity model. The lower bound c is specified by the researcher: $H = c > 0$.

First, from the item pairs for which $H_{ij} \geq c$, the pair is selected that has the highest H_{ij} value that also is significantly larger than 0. Significance is evaluated using the test statistic z (Molenaar et al., 1994). If none of the H_{ij}s satisfies these requirements, no scale can be formed. In each consecutive step an item f is selected and added to the already selected item set; the item selected (1) has a positive covariance with each of the already selected items; (2) has an H_f value with respect to the already selected items which is at least c; and (3) maximizes the common H coefficient of the already selected items together with item f across all possible choices from the remaining items. The selection of items for the first scale stops if all items have been selected or if none of the remaining items satisfies each of the conditions for selection into a scale. Because at each stage of the item selection process a large number of significance tests is performed, a progressive Bonferroni correction protects against chance capitalization across the selection steps.

Next, if possible, the item selection procedure selects items from the remaining items into a second scale using the same selection criteria. If there are items remaining, the procedure continues to select items into a third scale, and so on, until there are no more items remaining or until the items that remain do not satisfy the conditions for inclusion into a scale. The same test statistic z and the same Bonferroni correction are used

throughout the item selection procedure.

Factors That Affect Item Selection

Obviously, factors that influence H affect the outcome of the item selection procedure. The sensitivity of the selection procedure to the variance of the person parameter θ, while holding the δs and αs constant, was investigated for simulated unidimensional dichotomous data (Sijtsma & Prins, 1986). The two-parameter logistic model (Birnbaum, 1968) and a normal θ distribution were used to generate the data. All items were selected into one scale for all cases under consideration. When the variance of θ increased, the H value of the total scale increased. The H value of the scale decreased during the selection of the first 5 items, but did not decrease much more with the selection of more items.

If relatively weakly discriminating items were added to the item bank, these items were selected last. Their selection led to a sharp decrease in H. The δs of these items had no effect on this result. If one relatively strongly discriminating item was added to the item bank, this item was one of the two items selected first regardless of its δ.

For unidimensional polytomous data, Hemker & Sijtsma (1993) demonstrated that the result of the item selection procedure is hardly influenced by the number of answer categories. It is unknown, however, how multidimensionality of the data affects the outcome of the item selection. Suppose that in a test measuring d latent traits, each item measures only one trait. Different traits can have nonzero correlation, however. The desired result would be that (1) all items measuring the same trait are selected into one scale, and (2) items measuring different traits are selected into different scales. However, characteristics of the joint distribution of the d latent trait values $\theta_1, \theta_2, \ldots, \theta_d$, the αs of the items, and the spread of the item step difficulties will have an effect on the accuracy of the outcome. For example, holding everything else constant, the smaller the correlation between traits, the more items will be classified correctly into unidimensional scales. This will also depend on the lower bound c selected by the researcher. Note, however, that for a correlation larger than 0, items indirectly measure more than one trait: the trait for which the item was originally designed and the other trait(s) covarying with this trait.

If the αs are equal for all items in the multidimensional item bank, its effect on the outcome of the selection procedure will be relatively straightforward (Sijtsma & Prins, 1986). However, if the αs are unequal, results for a given lower bound will be more diffuse. Items may be rejected on the basis of low α. As a result, only relatively highly discriminating items measuring the same trait may be selected into one scale.

In this study, the correlation between traits, the mean α, and the variance of the αs across items measuring the same trait were varied to study their effects on the outcome of item selection. Other factors, such as the variance of the joint person parameter distributions and the spread of the δs, were held constant.

Lower Bound $H = c$ and Item Selection

The composition of the scales resulting from the selection procedure will be affected by the magnitude of the lower bound $H = c$. If $c = 0$ (the theoretically smallest lower bound given monotone homogeneity), the items are selected into scales for which $H_{ij} \geq 0$ for all item pairs and thus $H \geq 0$. Therefore, scales may have very low H values and may be practically useless. If $c = 1$ (the largest lower bound), the item selection procedure attempts to find a perfect Guttman (1944) scale and stops if none is available.

Very small lower bounds may yield one large scale that consists of items measuring different traits. Very large lower bounds may yield scales that do not include all items measuring the same trait, or they may yield more scales than there are traits. Between these two undesirable outcomes, a range of c values that yields the desired result hopefully can be located. For practical purposes, this range can be extended by allowing a suboptimal but practically acceptable outcome; for example, by requiring that (1) at least two-thirds of the items measuring the same trait are selected into one scale; (2) items measuring different traits

are always selected into different scales; and (3) the number of scales equals the number of traits. In this study, the range that leads to the perfect selection from the item bank (the correct range) and the practically acceptable ranges were determined as a function of the correlation between traits, the mean item α, and the variance of the αs across items measuring the same trait.

Because an analytic solution of the problem of finding the ranges of c values is difficult to achieve, and given the dependence on many characteristics of the items and the joint distribution of $\theta_1, \theta_2, \ldots, \theta_d$, a simulation study was conducted. From this study a practical strategy was formulated for the selection of c that results in the selection of the appropriate unidimensional scales from multidimensional data.

Lower Bound Values for Detecting Unidimensional Scales

Method

For d traits the number of correlations between trait pairs is $d(d-1)/2$; thus, many relationships could be studied. Here the simplest situation ($d = 2$) was investigated extensively; limited results were obtained for $d = 3$. From these results, it is plausible that the study of larger d will not lead to additional insights.

For the $d = 2$ case, the joint distribution of the latent traits θ_1 and θ_2 was standard normal. Six correlations between θ_1 and θ_2 were investigated: $\rho = 0.0, .2, .4, .6, .8, 1.0$. An item bank containing 18 items was used. Items 1 to 9 measured θ_1, and Items 10 to 18 measured θ_2. For each trait, the item numbering corresponded with increasing δ (i.e., Item 2 was more difficult than Item 1 and so forth). There were five answer categories per item because in many practical situations five-alternative Likert items are used.

Because the Mokken model of monotone homogeneity does not parametrically define ISRFs and because a parametric definition in combination with the joint distribution of θ_1 and θ_2 is necessary to generate datasets, the graded response model (Masters, 1982; Samejima, 1969) was used (see Equation 6).

The mean (M) α, denoted by α_M, of the items measuring θ_1 was equal to that of the items measuring θ_2. Three levels were investigated: $\alpha_M = 1.0$, $\alpha_M = 1.5$, and $\alpha_M = 2.0$. With a bivariate standard normal distribution, these values resulted in low, medium, and high quality items, respectively. The spread of the αs within a unidimensional item set had two levels: constant (C; no spread) and varying (V; positive spread). For the C condition, $\alpha_i = \alpha_M$ for all items. For Condition V, $\alpha_i = \alpha_M + v$, with $v = -.5$ for three items, $v = 0$ for three other items, and $v = .5$ for the remaining three items. The v values were randomly distributed across the items. The random assignment resulted in $v = (0, -.5, .5, .5, 0, -.5, 0, -.5, .5)$ for the items measuring θ_1 ($i = 1, \ldots, 9$), and $v = (.5, 0, -.5, .5, -.5, .5, 0, 0, -.5)$ for the items measuring θ_2 ($i = 10, \ldots, 18$). Note that in Condition C a restrictive version of the monotone homogeneity model is obtained, which also satisfies the model of double monotonicity for polytomous items [see Molenaar (1986, in press) for further details]. Because of the interdependence between the item α, the spread of the δs, and the population variance, the latter two were not included as factors in the design.

For each trait, the 9 item δs were equidistant between -2 and 2. Thus, δs from different subsets were matched. The τ_{is}s were equidistant between -1.5 and 1.5 for all items.

The result was a completely crossed $6 \times 3 \times 2$ design, with 6 levels of correlation between the traits, 3 levels of mean α, and 2 levels of variation of the α parameters (C and V). For every cell in this design, four replications each consisting of 2,000 simulees were generated. A sample of 2,000 simulees was assumed to be sufficiently large to obtain stable results, and four replications were considered sufficient to evaluate this assumption.

For $d = 3$, the three traits θ_1, θ_2, and θ_3 had a joint standard normal distribution. The three correlations between the traits were equal: $\rho_{12} = \rho_{13} = \rho_{23} = \rho$. Analogous to $d = 2$, for $d = 3$ there were 9 items per trait. A $6 \times 3 \times 2$ design was investigated. For every cell in the design one dataset consisting of 2,000 simulees was generated analogous to the situation with $d = 2$. Note that one dataset was used because for $d = 3$ the stability results obtained for $d = 2$ were used.

With two traits and 9 items per trait the correct result is obtained when two scales, each containing all 9 items measuring one particular trait, are selected. Such an outcome will be denoted [2:9,9]. The practically acceptable result has two unidimensional scales containing at least 6 items each, but not 9 items in both, and is denoted [2: ≥ 6, ≥ 6]. For $d = 3$ the correct result, [3:9,9,9], and the practically acceptable result, [3:≥6,≥6,≥6], are defined analogously. If $\rho = 1.0$ the data are unidimensional and, therefore, one unidimensional scale that contains all items should be found in all cases.

The software package MSP (Mokken Scale Analysis for Polytomous Items; Molenaar et al., 1994) was used for all calculations. MSP performs item selection using the test statistic z and the Bonferroni correction discussed above. For this simulation study, however, this inferential framework had almost no practical effectiveness for the following reason. Given that the model of monotone homogeneity is true, the null hypothesis $H = 0$ is true if the ISRFs have zero discrimination or if θ has zero variance. These conditions were not represented in this simulation study. In addition, the authors' experience has shown that the test for $H = 0$ is useful if the sample size is smaller than approximately 500. In larger samples the null hypothesis is almost always rejected. This is the result of large power against the null hypothesis if the monotone homogeneity model is the true population model. Because the sample size was 2,000, the items had positive α, and θ had positive variance, statistical testing had no effect.

Pilot Study

In a pilot study, datasets containing 400 simulees from some of the cells of the design with $d = 2$ were analyzed. For example, $\alpha_M = 1.5$ for all 18 items (Condition C) and correlation between traits $\rho = .4$ yielded the following item selection results.

For $c = 0$, one two-dimensional scale was found containing all 18 items with $H = .27$. For all $c < .20$ this same result was found. However, for $c = .20$ the item selection procedure resulted in two unidimensional scales each containing 9 items and each with $H = .41$. For all $.20 \leq c \leq .39$ this [2:9,9] result was found. For $c = .40$ two scales also were found, but one of the scales contained 8 items instead of 9. This outcome thus belonged to the class [2:≥6,≥6]. For $c = .41$ three scales were found: one with 8, one with 7, and one with 2 items. This result is not practically acceptable because the number of unidimensional scales is incorrect. For values of $c > .42$, the item selection procedure yielded other results that were not practically acceptable. For example, for $c = .45$ four small unidimensional scales were found containing 2 to 4 items. For $c = .50$ only two 2-item scales were found.

Intervals of c Values

In general, it can be concluded that if $c = 0$ and $\rho > 0$, most or all items were selected into one scale because the only requirement is that $\sigma(X_i, X_j) > 0$ ($H_{ij} > 0$) for all item pairs. If $\rho = 0$ (independent traits), many negative sample covariances will occur between items measuring different traits and thus more than one scale is expected to result.

If c increases starting from 0, then depending on the parameter setup of the ISRFs and ρ, there exists a value of c, say c_S, that is the smallest c that results in two unidimensional scales. There are two possibilities of interest here. First, [2:9,9] will be found if all items measuring different traits have lower sample covariances than items measuring the same trait. Second, if [2:9,9] is not found, the practically acceptable result [2:≥6,≥6] may be found if both kinds of covariances show only small differences. The result [2:≥6,≥6] will be found in particular if the αs vary within a set of items measuring one trait. In extreme cases (e.g., ρ very large), not even [2:≥6,≥6] will be found. For the pilot study, [2:9,9] was found with $c_S = .20$.

If c increases further starting from c_S, then two possibilities are of interest. If [2:9,9] was obtained at c_S then there exists a larger c, say c_{LC}, that is the largest c yielding a correct result. Values larger than c_{LC} result in imperfect outcomes. Thus, the correct result is obtained between c_S and c_{LC}. If [2:≥6,≥6] was obtained

at c_S, then there exists a larger c, say c_{LPA}, which is the largest c yielding a practically acceptable result. Values larger than c_{LPA} result in outcomes worse than [2: ≥6, ≥6]. For the pilot study, this resulted in c_{LC} = .39 and c_{LPA} = .40. Note that in no cell is it possible that $c_{LPA} < c_{LC}$ because lower bounds larger than c_{LPA} cannot result in [2: 9, 9]. It is expected that the choice of c has a similar effect on item selection for $d = 3$.

Results

Two traits. Because c_S hardly varied across replications, Table 1 shows the mean values of c_S across four replications for $\rho < 1.0$. Note that $\rho = 1.0$ represents unidimensionality; thus, c_S cannot be determined. In most cases (exceptions marked by *) these c_S values resulted in [2: 9, 9]. The value of c_S increased with increasing correlation between the two traits. For $\rho > 0.0$, c_S also increased with increasing mean αs within Condition V or C. No value shown in Table 1 for a condition means that no value of c resulted in a correct or a practically acceptable result.

Table 1
Mean c_S Values for $d = 2$ Averaged Across 4 Replications and 5 Levels of ρ, With $N = 2{,}000$ for 3 Levels of α_M and With α Varying (V) or Constant (C) Over Items

	$\alpha_M = 1.0$		$\alpha_M = 1.5$		$\alpha_M = 2.0$	
ρ	V	C	V	C	V	C
0.0	0.00	0.00	0.00	.02	.01	.01
.2	.07	.07	.11	.10	.13	.14
.4	.16*	.11	.20	.19	.25	.26
.6	.23*	.16	.32*	.28	.36	.36
.8	—	.20	—	.35	.51*	.46

For constant αs (C), the selection procedure always resulted in [2: 9, 9]. Under the V condition, this result was only found when ρ was small or when the mean α was large. When [2: 9, 9] was not obtained in Condition V, the items that were not selected had relatively low αs. In Condition V with $\alpha_M = 1.0$ and $\alpha_M = 1.5$, respectively, and $\rho = .8$, two unidimensional scales also were found, but because at least one of these scales contained fewer than 6 items c_S is not given. If a lower bound c was used that was smaller than the c_S value reported in Table 1, one scale that contained items measuring both traits was obtained.

The correlation between the traits (ρ) had no effect on c_{LPA} and c_{LC}. Furthermore, c_{LPA} and c_{LC} hardly varied across replications. Therefore, Table 2 shows the mean values of c_{LPA} and c_{LC} across the five correlations (omitting $\rho = 1.0$) and the four replications per cell. These means are based on observations of c_{LPA} and c_{LC} that had standard deviations (SDs) ranging from .0076 to .0166 across cells, with a mean SD of .0108.

The values of c_{LPA} and c_{LC} increased with increasing mean αs. Furthermore, c_{LC} was larger in Condition C than in Condition V. However, c_{LPA} was larger in Condition V than in Condition C.

If c exceeded c_{LPA}, the results (not shown) also differed in Conditions V and C. In Condition V, two unidimensional scales were found if c exceeded c_{LPA} by less than .10, but at least one of these scales had fewer

Table 2
Mean c_{LC} and c_{LPA} Values for $d = 2$ Averaged Across 4 Replications and 5 Levels of ρ, With $N = 2{,}000$ for 3 Levels of α_M and With α Varying (V) or Constant (C) Over Items

Type	$\alpha_M = 1.0$		$\alpha_M = 1.5$		$\alpha_M = 2.0$	
of c	V	C	V	C	V	C
c_{LC}	.11	.21	.30	.39	.46	.53
c_{LPA}	.26	.22	.43	.40	.56	.54

than 6 items. If the lower bound c exceeded c_{LPA} by .10 or more, then no scales were found or the number of items per scale was only 2 or 3.

In Condition C, the average difference between c_{LC} and c_{LPA} was only .01 (Table 2). As c increased ($c > c_{LPA}$) (not shown), the number of small scales increased. For example, if for a particular c two items measuring the same trait were not selected in a larger unidimensional scale, and these two items had a mutual H_{ij} value that was larger than c, this resulted in an additional scale containing these two items. A further increase of c resulted in a decrease in the number of scales, because the lower bound c became larger than the H_{ij} values.

The range of lower bounds that resulted in two distinct unidimensional scales can be inferred from the results for c_S, c_{LC}, and c_{LPA}. The lower bound ranges for the different levels of ρ and α_M are shown in Figure 1 for Condition V and in Figure 2 for Condition C. The black part of each column shows the range of lower bounds that yielded the correct result, $[2: 9, 9]$; the white part of a column shows the range of lower bound values that yielded the practically acceptable result, $[2: \geq 6, \geq 6]$. An asterisk (*) in Figure 1 means that neither result was found.

The range of lower bounds increased with increasing αs and with decreasing correlation between traits. The range that resulted in two unidimensional scales was larger for varying (V) αs (Figure 1), but the range of lower bounds resulting in $[2: 9, 9]$ was larger for constant (C) αs (Figure 2).

For the special case in which $\rho = 1.0$ (unidimensionality), the correct result is obviously that all items are selected into one scale. For Condition C this result was found between $c = 0.00$ and $c = .21$ for $\alpha_M = 1.0$, between $c = 0.00$ and $c = .39$ for $\alpha_M = 1.5$, and between $c = 0.00$ and $c = .53$ for $\alpha_M = 2.0$. For larger lower bounds, one or a few small scales consisting of two or three items were found. For much larger c, no scales were found. For Condition V, all items were selected in one scale between $c = 0.00$ and $c = .10$ for $\alpha_M = 1.0$, between $c = 0.00$ and $c = .30$ for $\alpha_M = 1.5$, and between $c = 0.00$ and $c = .46$ for $\alpha_M = 2.0$. With increasing c, first one smaller scale was found because a number of items were rejected, next one or a few very small scales were found and finally, no scales were found.

Three traits. The results for $d = 3$ closely resembled the results for $d = 2$. For $\rho < 1.0$, the values of c_S per cell, averaged across ρ, α_M, and Conditions V and C were .015 larger than for $d = 2$, with a SD of .017. The largest difference was .05. The values of c_{LC} and c_{LPA} for $d = 3$, averaged over ρ, α_M, and Conditions V and C were both .004 smaller than for $d = 2$, with SDs of .009 and .010, respectively. The largest differences for c_{LC} and c_{LPA} with corresponding c_{LC} and c_{LPA} values for $d = 2$ were $-.02$ and $-.03$, respectively. The results for $\rho = 1.0$ with $d = 3$ were comparable with the results for $\rho = 1.0$ and $d = 2$.

Research Strategies and an Empirical Example

It is common practice to use this item selection procedure with only one lower bound value, very often $c = .3$ (Mokken & Lewis, 1982). This value was proposed to obtain a sufficiently accurate ordering of persons. However, this study has shown that there is not a unique lower bound or even a unique range of lower bounds that will indicate whether the data are multidimensional. Such ranges vary across the particular choice of ISRFs and the joint distribution of $\theta_1, \theta_2, \ldots, \theta_d$. Thus, the use of one c value may be not enough to decide whether an item set is multidimensional.

A relatively simple way to obtain information about the multidimensionality of the data is to implement the item selection procedure several times, starting with $c = 0.0$ or $c = .05$. These values will reveal whether there are scales with correlations close to 0. These scales should not be joined. Next, the item selection procedure can be implemented for $c = .30$, $c = .40$, and $c = .50$.

More detailed information can be obtained by implementing the item selection procedure more often, starting with $c = 0.0$ and in each consecutive step increasing c with, say .05, until c is approximately .55. The precision of the results can be manipulated by selecting smaller or larger increases of c. Following this

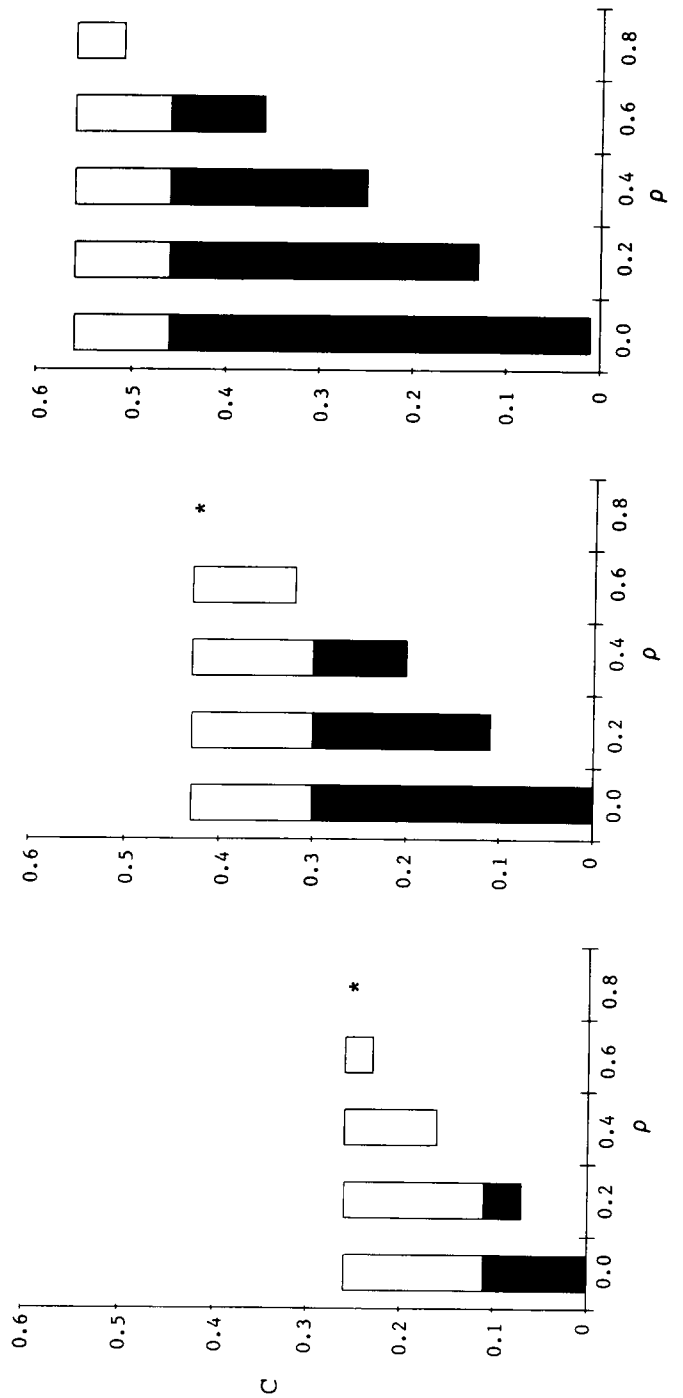

Figure 1
Lower Bound Ranges That Resulted in Two Distinct Unidimensional Scales for $d = 2$, Averaged Across Four Replications for 5 Levels of ρ and 3 Levels of α_M, for Condition V

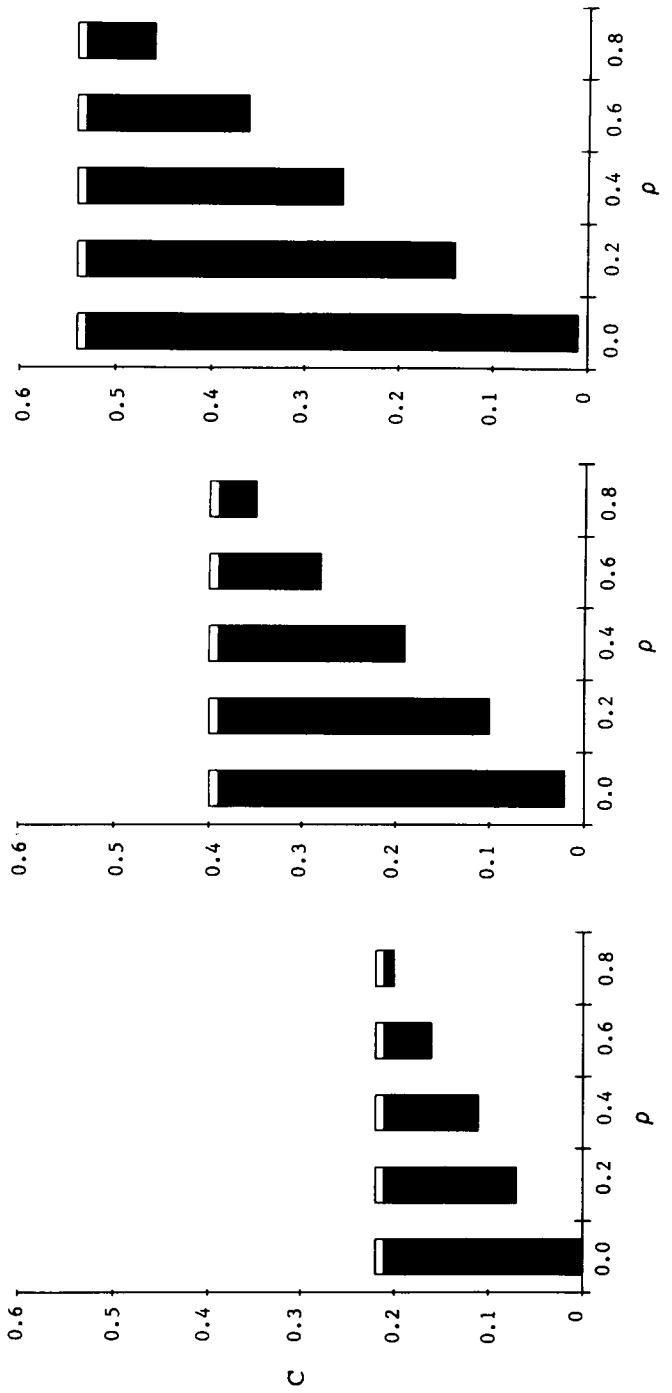

Figure 2
Lower Bound Ranges That Resulted in Two Distinct Unidimensional Scales for $d = 2$, Averaged Across Four Replications for 5 Levels of ρ and 3 Levels of α_M, for Condition C

strategy means running the item selection procedure 12 times. However, for an experienced user of MSP this will take very little time.

The results from this simulation study suggest how these strategies be used with empirical datasets. Two cases should be distinguished: multidimensionality and unidimensionality.

The typical pattern of results with multidimensional data for varying lower bound c is that with increasing c the following stages can be observed: (1) most or all items are in one scale; (2) two or more unidimensional scales are formed; and (3) two or more smaller scales are formed and several items are rejected. This study indicates that the results from the second stage be taken as final. Thus, the scales found in that stage may be used as separate unidimensional scales.

With unidimensionality, the typical pattern of results with increasing c is: (1) most or all items are in one scale; (2) one smaller scale is found; and (3) one or a few small scales are found and several items are rejected. If this pattern of results is found with an empirical dataset, consider the results from the first stage as final. For practical purposes, test length, the value of H, and the reliability of scale scores also should be taken into consideration with either of the outcomes pertaining to unidimensionality and multidimensionality. The main difference between the stages observed for multidimensionality and those for unidimensionality is that for multidimensionality the scale found in Stage 1 splits into two or more scales whereas for unidimensionality this scale mainly remains intact.

Example Application

The item selection procedure was applied to empirical data from an investigation of annoyance due to industrial malodors (Cavalini, 1992). The questionnaire consisted of 17 four-category items administered to 828 respondents.

Factor analysis (Cavalini, 1992, pp. 53–54) revealed several solutions; however, the most interpretable had four factors [4: 7, 4, 3, 3]. Scale 1 measured a mixture of an emotional and an avoidance reaction and contained 7 items (Items 3, 6, 8, 13–16); Scale 2 measured the rational effort to do something about the malodor problem and contained 4 items (Items 5, 7, 9, 11); Scale 3 measured the effort to save the laundry from the bad outside air and contained 3 items (Items 1, 2, 4); and Scale 4 measured the emotional acceptance of the situation and contained 3 items (Items 10, 12, 17) (Cavalini, p. 53). The correlations among the four scales ranged from –.31 to .47. Other than one high loading, each item had small loadings on the other factors.

MSP and the methodology presented above were used in order to investigate whether a solution that had the same interpretation could be obtained. Starting with $c = 0.00$, with each consecutive step c was raised by .05 until .55. Table 3 shows the predicted pattern of results, starting with most items in the same scale and ending with a few small scales and most items rejected. For $c = 0.00$, 14 items were in one scale and 3 items were in another scale. Between $c = .20$ and $c = .40$, first three and then (at $c = .30$) four scales were formed. The same 3 items (Items 10, 12, 17) that formed one scale for $c = 0.00$ constituted this scale until $c = .30$.

These results suggest that either three or four scales should be accepted as the end result. Here considerations concerning the number of items per scale, the reliability per scale score, and the interpretation of the meaning of the scales can be used for a final decision. Note that the solutions with three (at $c = .25$) and four (e.g., at $c = .35$) scales both contained Scale 3 and Scale 4 from Cavalini's (1992, pp. 53–54) factor analysis. The three-scale solution also had a union of most items from Scale 1 and all items from Scale 2. This seems somewhat unfortunate, given that the first scale had a strong emotional component whereas the second reflected a more rational attitude. The four-scale solution basically had the same four scales in terms of interpretation as the factor analysis solution. Because the questionnaire was used for research but not for individual diagnosis, reliability is less important and interpretation will be more important.

Table 3
Scales Determined Using the Mokken Item Selection Procedure

		Item Numbers				
c	Result	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
0.00	2:14,3	1–9,11,13–16	10,12,17			
.05	2:14,3	1–9,11,13–16	10,12,17			
.10	2:14,3	1–9,11,13–16	10,12,17			
.15	2:14,3	1–9,11,13–16	10,12,17			
.20	3:12,3,2	1–7,9,11,13–15	10,12,17	8,16		
.25	3:9,3,3	3,5–7,9,11,13–15	10,12,17	1,2,4		
.30	4:8,3,3,2	3,5–7,9,11,13,14	10,12,17	1,2,4	8,15	
.35	4:5,2,3,4	3,5,7,9,11	10,17	1,2,4	6,13,14,15	
.40	4:4,2,2,2	5,7,9,11	10,17	2,4	13,14	
.45	4:4,2,2,2	5,7,9,11	10,17	2,4	13,14	
.50	5:2,2,2,2,2	7,9	5,11	2,4	13,14	8,15
.55	2:2,2	7,9	2,4			

Discussion

The simulation study had several limitations. First, the maximum number of dimensions was three. This covered most, but not all, relevant situations. However, there is no reason to believe that results would be much different for more than three dimensions. Second, each item measured one dimension. In practice, it is reasonable to assume that items can be multidimensional (e.g., an arithmetic item can also require verbal skills). However, because of correlations between the underlying traits such items will be positively correlated. Third, dimensions were represented by equal numbers of items and this number was not varied across the design. Equal numbers of items reflects an effort to have subtest scores of approximately equal reliability. However, in practice, the number of items may be different. Furthermore, the spread of the item locations was fixed. But it was argued that manipulating the discriminations would have effects comparable to manipulating item locations. The factors that were manipulated here were the most informative given that almost nothing was known about item selection from multidimensional item banks in the framework of the Mokken approach to scaling.

Note that the research strategies are based on an admittedly limited but well-chosen and completely crossed design. Empirical datasets often will have relatively irregular characteristics that may lead to deviations from the patterns predicted here. However, the design studied here is believed to be appropriate for many practical datasets.

An interesting topic for future research would be further comparison of the methodology proposed here with results from other methods such as factor analysis. For example, the characteristics of items or item sets that produce different results may be investigated, as in the empirical example presented here. Such research should preferably use empirical data in addition to simulated data, as in this example. This future research should also address the limitations of the present simulation study.

The strategy proposed here can be used to investigate whether an item set is multidimensional. Note, however, that finding the correct number of unidimensional scales is often not the only goal in scale construction. Perhaps a scale should not be split into two unidimensional scales measuring highly correlating traits. In that case, reliability and validity will be important issues.

References

Batley, R. M., & Boss, M. W. (1993). The effects on parameter estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement,*

17, 131–141.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 396–479). Reading MA: Addison-Wesley.

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries.* Unpublished doctoral dissertation, University of Groningen, The Netherlands.

de Gruijter, D. N. M. (1993). Comparison of the nonparametric Mokken model and parametric IRT models using latent class analysis. *Applied Psychological Measurement, 18,* 27–34.

Ellis, J. L., & van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika, 58,* 417–429.

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

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton NJ: Princeton University Press.

Hemker, B. T., & Sijtsma, K. (1993). A practical comparison between the weighted and the unweighted scalability coefficient of the Mokken model. *Kwantitatieve Methoden, 14*(44), 59–73.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21,* 1359–1378.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45,* 507–530.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

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

Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement, 19,* 323–335.

Meijer, R. R., Sijtsma, K., & Smid, N. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

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

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

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement, 10,* 279–285.

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

Molenaar, I. W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën [An exercise in item response theory for three ordered response categories]. In G. F. Pikkemaat & J. J. A. Moors (Eds.), *Liber amicorum Jaap Muilwijk* (pp. 39–57). Groningen: Econometrisch Instituut.

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

Molenaar, I. W. (in press). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern psychometrics* (pp. 361–373). New York: Springer.

Molenaar, I. W., Debets, P., Sijtsma, K., & Hemker, B. T. (1994). *User's manual for the computer program MSP* (Ver. 3.0). Groningen: iec ProGAMMA, Rijksuniversiteit Groningen.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361–373.

Rosenbaum, P. R. (1987). Comparing item characteristic curves. *Psychometrika, 52,* 217–233.

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

Sijtsma, K., & Prins, P. M. (1986). Itemselectie in het Mokken model [Item selection in the Mokken model]. *Tijdschrift voor Onderwijsresearch, 11,* 121–129.

Stout, W. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 49,* 293–325.

Wood, R. (1978). Fitting the Rasch model—A heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

Author's Address

Send requests for reprints or further information to Bas T. Hemker, Utrecht University, Department of Methodology and Statistics, Faculty of Social Sciences, P. O. Box 80.140, 3508 TC Utrecht, The Netherlands.