Assessing Dimensionality by Maximizing *H* Coefficient–Based Objective Functions

Alexandra A. H. van Abswoude, Maastricht University, Netherlands
Jeroen K. Vermunt, Tilburg University, Netherlands
Bas T. Hemker, Citogroep, Netherlands

Mokken scale analysis can be used for scaling under nonparametric item response theory models. The results may, however, not reflect the underlying dimensionality of data. Various features of Mokken scale analysis—the *H* coefficient, Mokken scale conditions, and algorithms—may explain this result. In this article, three new *H*-based objective functions with slight reformulations of Mokken scale analysis in the unidimensional and multidimensional cases are introduced. Deterministic and stochastic nonhierarchical clustering algorithms reduced the probability of obtaining suboptimal solutions. A simulation study investigated whether these methods can determine the dimensionality structure of data sets that vary with respect to item discrimination, item difficulty, number of items per trait, and numbers of observations per test. Furthermore, it was investigated whether deterministic and stochastic algorithms can generate approximately global optimal solutions. The method based on the average within-scale H_i combined with a stochastic nonhierarchical clustering algorithm was the most successful in dimensionality assessment. *Index terms: optimizations, multidimensionality, nonparametric item response theory, sequential clustering, scaling, stochastic algorithms*

A mathematics test that draws on the student's spatial insight, calculus, and arithmetic abilities is known as a multidimensional test. In general, a test or questionnaire that is sensitive to more than one trait, ability, or characteristic is denoted as multidimensional. One type of multidimensionality involves the situation where the item pool is sensitive to multiple latent traits and the items are sensitive to one dominant latent trait. For an unequivocal interpretation in those measurement situations, it is convenient to select items into multiple sets, each sensitive to one dominant trait, and develop scale scores for each trait separately.

Four nonparametric item response theory (IRT) methods can be used to select one or more sets of items sensitive to a single trait from a multidimensional data matrix. Dedicated software packages are DETECT (Kim, 1994; Zhang & Stout, 1999), DIMTEST (Nandakumar & Stout, 1993; Stout, Goodwin Froelich, & Gao, 2001), HCA/CCPROX (Roussos, Stout, & Marden, 1998), and MSP (Mokken, 1971; Molenaar & Sijtsma, 2000; Sijtsma & Molenaar, 2002). These methods all use observable consequences of the monotone homogeneity model (MHM; Mokken, 1971). A set of items that satisfies the MHM is unidimensional (i.e., sensitive to a single latent trait), is locally independent (i.e., the item responses are statistically independent given a fixed value of the latent trait), and meets the monotonicity assumption (i.e., the probability of answering an item correctly is an increasing function of the latent trait; e.g., Holland & Rosenbaum, 1986).

Applied Psychological Measurement, Vol. 31 No. 4, July 2007, 308–330
DOI: 10.1177/0146621606295196

The methods vary in their focus on each of the particular MHM assumptions. A relaxation of nonparametric IRT's local independence assumption, denoted as weak LI (e.g., McDonald, 1985; Stout, 1987), is used to evaluate the relationship between item pairs in DETECT, DIMTEST, and HCA/CCPROX. In DETECT and HCA/CCPROX, the items are partitioned into clusters in such a way that locally independent sets of items are obtained as much as possible, and in DIMTEST, weak LI of the responses is tested. Although multidimensionality may not be the only possible explanation that weak LI does not hold, this approach appears to be a rather direct one to assess dimensionality (van Abswoude, Van der Ark, & Sijtsma, 2004). The methods have a few disadvantages; that is, these methods are not suitable for tests having few items and a modest sample size (sample sizes of 2,000 and less are regarded as being small; Stout, 1987); they are sensitive to the strength of each scale (i.e., different numbers of items or different discrimination of the items between scales; see van Abswoude et al., 2004); and their statistics have some bias (see Roussos & Ozbek, 2003; Stout et al., 2001; Zhang, Yu, & Nandakumar, 2003).

The focus of MSP's method, Mokken scale analysis (MSA), is on creating scales rather than on dimensionality assessment. Items joined into a scale using MSA satisfy an observable consequence of the MHM on one hand, and a user-defined condition on the other hand. The user-defined condition allows one to choose the minimal discrimination power of items in a scale. The strength of the relationship between item scores is quantified by means of the H coefficient (Loevinger, 1948; Mokken, 1971), a normed covariance that corrects for the maximum covariance possible given the marginal distribution of the items. This coefficient need not be calculated for each latent trait value and thus requires fewer subjects and fewer items than the conditional statistics used in the local independence-based methods. A disadvantage is that its sequential scaling algorithm may not yield the best possible solution. Technically, the best solution for one scale is the item set having the highest H value for as many items as possible and satisfying the scale conditions. Practically, this disadvantage could mean that a scale has less strength to measure the underlying trait and/or consists of fewer items than if an alternative algorithm were used. For a theoretical comparison of the four methods, see van Abswoude et al. (2004).

The main purpose of this article is to implement a new algorithm in MSA that allows us to keep the general focus of the method intact but resolves the problems associated with the old algorithm. We use stochastic and deterministic versions of a nonhierarchical clustering algorithm for this purpose. New objective functions are introduced in which MSA is adapted to these new algorithms. These new objective functions define the problem of finding more than one scale in a slightly different way than the sequential MSA method. The necessity of this redefinition and its consequences for scaling results are discussed.

Although we intend to keep the scaling focus intact, in this article, the scale conditions are generally ignored. The main reason for this is simplicity: Before we add restrictions to the problem, we first want to investigate how well the new functions work. A consequence of adopting this approach is that the new method is investigated as a dimensionality assessment tool. This has a number of advantages. If the scale conditions are not incorporated, weakly discriminating items, for which the assignment of items into scales is the most difficult, can be selected into scales, and thus the limitations of the method can be investigated. Furthermore, we can find out whether the new function, the scale conditions, or the algorithm is responsible for splitting or joining of item pairs into clusters. Suggestions for how to extend the new MSA method with the scale conditions are discussed elsewhere (van Abswoude, 2004).

Using a simulation study, we determined how successful these methods are in finding the underlying dimensionality of a data matrix, which answers the first research question. In particular, we investigated the correspondence between the solution (i.e., the obtained sets of items or clusters) that maximized the object function concerned and the true dimensionality. The second research question

is connected to the preferred algorithm. We judged the success of each algorithm by the number of times it yielded a (near) global solution and by the number of iterations it needed to converge.

MSA

MSA uses Loevinger's H coefficient as a scalability coefficient (Loevinger, 1948; Mokken, 1971). The H coefficient can be expressed in terms of Guttman errors (Guttman, 1950). Let $\mathbf{X} = (X_1, \ldots, X_I)$ be the vector of I binary scored item response variables (items), and let $x = (x_1, \ldots, x_I)$ be their realizations (i.e., 0 denotes incorrect; 1 correct). In addition, let π_i denote the proportion of subjects answering item i correctly, and let all items be ordered such that $\pi_i \geq \pi_j$. Then, a subject answering an easy item i incorrectly and a difficult item j correctly produces a Guttman error. A larger number of Guttman errors than expected under the MHM in combination with the distribution of the latent trait(s) may be due to misfit of one or a few subjects (person misfit; e.g., Emons, 2003; Karabatsos, 2003; Meijer, 1994), the misfit of one or two items in specific subgroups of subjects (item bias), or the misfit of one or more items driven by unintended latent variables (multidimensionality; e.g., Stout, 1987; van Abswoude et al., 2004). Person misfit and item bias can in fact be seen as special cases of multidimensionality, in which instead of an extra unobservable trait, a grouping variable either distinguishes one subject from the rest or distinguishes different subgroups.

Let F_{ij} denote the observed number of Guttman errors, and $E_{ij} = N\pi_j(1 - \pi_i)$ the expected number of Guttman errors under marginal independence. The H coefficient for an item pair (i, j) is defined as

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{1}$$

One may note that $H_{ij} = 0$ when items i and j show exactly as many Guttman errors as expected under marginal independence, and $H_{ij} = 1$ when no Guttman errors are observed. The H coefficient can also be written as

$$H_{ij} = \frac{\text{cov}(X_i, X_j)}{\text{cov}(X_i, X_j)_{\max}} \tag{2}$$

(Loevinger, 1948; Mokken, 1971). The H coefficient of a single item i in a scale consisting of I items equals

$$H_i = \frac{\sum_{j \neq i} \text{cov}(X_i, X_j)}{\sum_{j \neq i} \text{cov}(X_i, X_j)_{\max}}. \tag{3}$$

Also, the overall H coefficient for a set of items can be written as a normed covariance; that is,

$$H = \frac{\sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \text{cov}(X_i, X_j)}{\sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \text{cov}(X_i, X_j)_{\max}}. \tag{4}$$

Alternatively, we could define the scale H as a weighted sum of the items H_is or the bivariate H_{ij}s (Mokken, 1971), respectively:

$$H\sum_{i1}^{I1}\sum_{ji1}^{I} E_{ij}H_i\sum_{i1}^{I1}\sum_{ji1}^{I} E_{ij} \qquad \sum_{i1}^{I1}\sum_{ji1}^{I} E_{ij}H_{ij}\sum_{i1}^{I1}\sum_{ji1}^{I} E_{ij}. \tag{5}$$

Mokken (1971, pp. 149-152) showed that the MHM implies that $0 \leq H_{ij} \leq 1$, $0 \leq H_i \leq 1$, and $0 \leq H \leq 1$. Thus, positive values of these coefficients are necessary for the MHM to hold. The relationship between H_{ij}, H_i, and H is the following: $\min(H_{ij}) \leq \min(H_i) \leq \max(H_i) \leq \max(H_{ij})$ (e.g., Hemker, Sijtsma, & Molenaar, 1995; Mokken, 1971). We restrict our attention to dichotomously scored items. The generalization of our methods to polytomous items (using Equations 2, 3, and 4) is straightforward (e.g., Hemker et al., 1995; Molenaar, 1991).

Theoretically, a Mokken (1971) scale is defined as

Condition 1: $\text{cov}(X_i, X_j) > 0$, for all $i \neq j$, and
Condition 2: $H_i \geq c$, for all i, where c is a user-defined constant between 0 and 1 (default, $c = .30$).

The first scaling condition, which can be restated as $H_{ij} > 0$, is necessary but not sufficient for the MHM to hold (Mokken, 1971 also see Holland & Rosenbaum, 1986). The second scaling condition serves a practical purpose and allows the user to manipulate the minimum discrimination of items joined into scales. Given the choice of c, not all items may be scalable. The scalable items agreeing with the MHM do, however, contribute to the correct ordering of subjects on the latent variable measured by each scale (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997). For the interpretation of the strength of a scale, Mokken (1971, p. 185) derived the following rules of thumb: $.30 \leq H < .40$ constitutes a weak scale; $.40 \leq H < .50$ a medium scale; and $H \geq .50$ a strong scale. Mokken considered $c = .30$ a reasonable minimal requirement for item quality. The appropriate value of c depends on the researcher's purpose of scaling. When highly scalable (or high discrimination) items are required, c needs to be high. Coefficient H also tends to be higher when the dispersion of items is larger (Roskam, Van den Wollenberg, & Jansen, 1986). For more information on the effect of c on dimensionality results, see Hemker et al. (1995), Molenaar and Sijtsma (2000), and van Abswoude et al. (2004).

Having a set of items with high H coefficients (Equation 2 or 3) does not necessarily imply that the set is sensitive to a single latent trait. For example, when traits are moderately correlated, despite multidimensionality, H can be high (van Abswoude et al., 2004). On the other hand, "Hemker et al. (1995) showed for polytomous items that the dominant dimensions of a data matrix may be found when various values for c are used. In fact, because the scale conditions are necessary but not sufficient for satisfying the MHM, one should check model assumptions. Let R_{-i} denote the total score on a set of items minus the score on item i. The program MSP then provides a tool to check the monotonicity of each item via nondecreasing $P[X_i = 1|R_{-i}]$ in R_{-i}, known as manifest monotonicity (Junker, 1993). Methods such as DIMTEST could be used in addition to MSP to ascertain that Mokken scales satisfy weak LI.

MSP uses a sequential clustering algorithm to select items into scales. Sequential item selection as defined by Mokken (1971) and as incorporated in the MSP method has the following stepwise

procedure. Item selection starts by joining of the item pair (i, j) with the highest H_{ij} under the restriction that it is significantly positive. This is the start set of the procedure. Then, out of all remaining items, the item i that yields the highest H with the already selected items is added to the start set under the following three restrictions: First, item i should have a positive covariance with each of the already selected items (Condition 1); second, H_i with respect to the already selected items should be significantly positive; and third, the H_i with respect to the already selected items should satisfy $H_i \geq c$ (Condition 2). This step is repeated until no item that satisfies the scaling criteria remains. Once this occurs, the first Mokken scale has been formed. If, after forming a scale, more than one item remains, the procedure is repeated to form a second, and a third (and so on) Mokken scale. Details about significance testing, the treatment of ties, or other aspects of the sequential algorithm can be found in Mokken (1971) and Molenaar and Sijtsma (2000).

Typical solutions found with this sequential procedure will be illustrated by means of a small example using simulated data. Assume we have data on a linguistics test having items on three topics: grammar (θ_1, 20 items), meaning (θ_2, 10 items), and punctuation (θ_3, 10 items). For such a test, one can easily imagine that the underlying abilities are correlated: We used $r(\theta_1, \theta_2) = .4$, $r(\theta_1, \theta_3) = .2$, and $r(\theta_2, \theta_3) = .2$. In addition, items may be sensitive to more than one trait. Then, starting out with the best item pair, two grammar items, using the sequential MSA method the following partitioning is found:[1] 25 items in Scale 1 (θ_1, θ_2; $H = .51$), 5 items in Scale 2 (θ_2; $H = .46$), and 10 items in Scale 3 (θ_3; $H = .60$). If we deviate from the default setting of MSP and use the next best pair as the starting set (i.e., two meaning items), we find 10 items in Scale 1 (θ_2; $H = .53$), 20 items in Scale 2 (θ_1; $H = .60$), and 10 items in Scale 3 (θ_3; $H = .63$). The combination of dependence on the start set and the inability to move items into better fitting clusters are the drawbacks of the sequential method in a nutshell (see Molenaar and Sijtsma, 2000, for instructions on how to cope with these issues in the current program MSP).

Problems such as this one can be expected to occur in scaling contexts where the underlying traits are significantly correlated (say, larger than .40; van Abswoude et al., 2004) or where items load on more than one trait. These conditions are typical for many test data situations.

Alternative Clustering Methods

In this section, we introduce three new methods that might improve MSA's optimization in a single as well as in a multiple scale context. To evaluate the quality of a partitioning consisting of items that are joined into one or more clusters, we need functions based on H. Each new function is called an objective function, and its purpose is to find a scaling solution that maximizes its value such that, for example, the highest value of the H coefficient for all clusters is obtained simultaneously. The proposed algorithms for maximizing the objective functions allow single items to be moved to a cluster where the fit may be better and allow multiple clusters to be formed simultaneously. In the following sections, we introduce three new objective functions based on H_{ij}, H_i, and H, respectively, and explain how these objective functions can be maximized.

Objective Functions Using the H Coefficient

Objective function O_1 was inspired by the work of Kim (1994) and Zhang and Stout (1999). $\eta_{ij} = 1$ if items i and j are in the same cluster k in partitioning P, and $\eta_{ij} = -1$ otherwise. Then, O_1 is defined as

$$O_1 = \frac{2}{I(I-1)} \sum_{1 \leq i < j \leq I} \eta_{ij}(H_{ij} - c^*). \tag{6}$$

The idea behind this objective function is that a good partitioning is achieved when pairs of items with high H_{ij}s are joined in the same cluster, and pairs with low H_{ij}s are put in different clusters. The partitioning that maximizes the objective function is referred to as P*. The multiplication by η_{ij} is incorporated in Equation 6 to encourage item pairs with a high H_{ij} to be joined into the same cluster and item pairs with a low or a negative H_{ij} to be put into different clusters. This is because the contribution of pair (i, j) to O_1 is positive if $H_{ij} - c^* > 0$ and $\eta_{ij} = 1$, and if $H_{ij} - c^* < 0$ and $\eta_{ij} = -1$. The contribution to O_1 is negative if $H_{ij} - c^* > 0$ and $\eta_{ij} = -1$, and if $H_{ij} - c^* < 0$ and $\eta_{ij} = 1$. Variations of O_1 can be obtained not only by different choices of c^*, but also by changing the definition of η_{ij}. For example, by setting $\eta_{ij} = 0$ for items i and j, which are not in the same cluster k, and $\eta_{ij} = 1$ otherwise, the objective function would target only the within-cluster scalability.

Rewriting Equation 6 makes the effect of c^* on the final clustering solutions clearer:

$$O_1 = \frac{2}{I(I-1)} \sum_{1 \le i < j \le I} \eta_{ij} H_{ij} - \frac{2}{I(I-1)} \sum_{1 \le i < j \le I} \eta_{ij} c^*. \tag{7}$$

If $c^* = 0$, the sum most right of Equation 7 equals zero. Thus, P* is found when all items, except those having many negative H_{ij}s with other items, are joined in one large set. With $c^* = 1$, which expresses the maximum value of H_{ij}, the rightmost sum of Equation 7 is maximized when items are distributed equally across the K clusters. The objective function is maximized with K equally large sets consisting of item pairs that jointly yield the highest ΣH_{ij} (see Equation 7), which means that $c^* = 1$ is appropriate only when item sets are equal in size. Because researchers do not know the latent trait composition of their test, $c^* = 1$ is useless in practice. Choosing a fixed value for c^*, for example, $c^* = .3$, is not advisable either because the suitability of a c^* value may depend on the properties of the investigated items. An alternative is to derive c^* from the data: $c^* = 2/I(I-1)\Sigma_{i \neq j} H_{ij}$, that is, the average H_{ij} of all item pairs (denoted $\bar{H}_{ij}$). One can easily verify that $O_1 = 0$ when $c^* = \bar{H}_{ij}$ and all items are entered in a single cluster. This provides a benchmark against which we can compare other solutions.

We investigated the appropriateness of the four proposed c^* values (0, .3, 1, and H_{ij}) for simulated data with two moderately correlated traits (i.e., .4). The algorithms and the model used for generating data are the same as in the main simulation study. Item discrimination (hi = high discrimination, or mo = moderate discrimination) and the length of the two subtests (15 items or 30 items) were varied. The results presented in Table 1 show that for $c^* = 0$ all items were joined into one cluster, regardless of the type of data. Using $c^* = .3$ yields the simulated structure for moderate discrimination items but not for high-discrimination items. This is because $H_{ij} > .3$ for most item pairs in the high-item discrimination condition, and as a result items sensitive to different traits are joined into one cluster. Using $c^* = 1$ works only for data having equal numbers of items for each trait (i.e., the total pool is split correctly), not for unequal numbers. Using $c^* = \bar{H}_{ij}$ yields the simulated structure for equal numbers of items and has one misclassified item for unequal numbers. We investigated the use of $c^* = \bar{H}_{ij}$ in O_1 more extensively in the main study.

The second objective function, denoted as O_2, can be interpreted as the average H_i within clusters of a partition. The objective function is used to maximize the item scalability. Before defining O_2, some additional notation is needed. Let k again denote an arbitrary cluster of items ($k = 1, \ldots, K$), and let H_i^k be the H_i value of item i with respect to the other items in cluster k. Let $\eta_i^k = 1$ if $i \in k$ at P (i.e., when item i is in cluster k), and $\eta_i^k = 0$ otherwise. The second objective function for evaluating a K-cluster partitioning P is

$$O_2 = I^{-1} \sum_{i=1}^{I} \sum_{k=1}^{K} \eta_i^k H_i^k. \tag{8}$$

Table 1
Effect of Using Different c^* Values of O_1 on Obtained Dimensionality Results

		c^*			
Test Composition	Discr.	0	.3	1	$\bar{H}_{ij}$
15/15	Hi	[30]	[30]	True	True
	Mo	[30]	True	True	True
15/30	Mo	[45]	[18/27]	[21/24]	[16/29]

Note. Discr. = Discrimination; hi = high discrimination; mo = moderate discrimination; true = the simulated structure was obtained; otherwise, the number of items obtained in each cluster is presented in brackets.

We use normalizing constant I^{-1} and indicator η_i^k to make O_2 easily interpretable as the average H_i within clusters. Note that with O_2 all elements of a partitioning are evaluated and not just one element at a time as in the sequential MSP method. This means that O_2 can be used to search for that partitioning of items that produces the highest H_i for all items. This property may resolve the problems discussed for MSP. As one can observe, we do not specify a constant c (Condition 2) or any related constant such as c^* in O_1. One may further note that maximizing O_2 may not yield the same solution as maximizing H. We use O_2 because of its direct relationship to the second Mokken scaling condition.

Let H^k denote the overall H for cluster k. The third objective function, denoted as O_3, equals the average within-cluster H:

$$O_3 = K^{-1} \sum_{k=1}^{K} H^k. \tag{9}$$

Out of the three presented objective functions, O_3 is most similar to MSP's original objective function.

The three objective functions presented above have in common that they make use of an average of the pairwise H_{ij}s. A difference is that O_1 uses the arithmetic mean of the H_{ij}s, whereas O_2 and O_3 use weighted normalized sums of the H_{ij}s. Another difference is that O_1 targets both within-cluster similarities and between-cluster differences, whereas O_2 and O_3 target only within-cluster similarities.

NHCA Algorithm

A well-known nonhierarchical clustering analysis (NHCA) algorithm is used to optimize O_1, O_2, and O_3. It is similar to the K-means algorithm (e.g., Berthold & Hand, 1999). Let t $(t = 1, \ldots, T)$ represent the iteration number.

In an NHCA algorithm, first an initial configuration is constructed. This means that each item i is assigned to its initial cluster k. At iteration t, the quality of the partitioning is evaluated using O_1, O_2, or O_3, and one item i is moved from a cluster k to another cluster k'. These steps are repeated until the process has converged.

The NHCA algorithm can be implemented in different ways. In our implementation, one item is moved at a time, but it would also have been possible to move more than one item per iteration. We choose not to move multiple items per iteration because that turned out to yield much less stable algorithms.

When applying the algorithm, we aim to find the partitioning that yields the highest value of the objective function given all possible partitionings. One may note that the objective function is

a discrete function whose value depends on the partitioning. As explained earlier, the solution space is bounded only by the selected value for K (i.e., no Mokken scale conditions are imposed). The best solution is the one that maximizes the objective function. This solution is known as the global optimum solution. Frequently, however, this objective function is multimodal, meaning that there are several local maxima in addition to a global maximum. In this article, the highest maximum that is obtained by running each of the algorithms is denoted the global maximum. Strictly speaking, this solution is globally optimal only by approximation because not all possible solutions were investigated.

In general, because simple deterministic algorithms have a higher likelihood of yielding a local maximum, we used stochastic algorithms in addition to deterministic ones. We chose to use an NHCA algorithm instead of simplex- or evolutionary-type algorithms (e.g., Danzig & Thapa, 1997; Michalewicz, 1996) because it closely resembles MSP's original algorithm and because it has the potential to resolve the optimization problems discussed for the sequential approach. Another attractive property is that deterministic and stochastic elements can easily be incorporated. These elements influence the likelihood that a global solution will be obtained and the speed of the algorithm. Deterministic and stochastic elements can be introduced into the NHCA at two occasions: at the initial configuration and at the move of a single item into a different cluster.

Initial Configuration

In the random initial configuration condition, items are randomly assigned to one of K clusters with equal probability (i.e., we use the default random number generator of Borland Pascal, Version 7.0). A random initial configuration requires no additional information and thus is simpler than its deterministic counterpart. Methods that use a random-start configuration may be repeated several times so that some may yield global optimal solutions (e.g, Michalewicz, 1996).

In the nonrandom initial configuration condition, one may use the K-cluster partitioning based on a priori knowledge. Another option is to start with the partitioning obtained with another clustering method, such as a sequential clustering or a hierarchical clustering procedure. Here, we used the partitioning obtained with the sequential clustering without testing procedure of MSP as the starting point. This nonrandom-start configuration can be expected to be close to the underlying dimensionality, and it can therefore be expected that few iterations will be needed to arrive at a final solution.

A practical complication with this start configuration is that sequential clustering may yield a larger number of clusters than NHCA and that some items may not be selected at all. There are different ways to remedy this, such as trying to obtain a K-cluster solution by manipulating the constant c in sequential clustering. This may, however, be rather cumbersome. In the simulation study, we therefore took a shortcut and randomly redistributed items from extra clusters and nonselected items over the K available clusters to get the desired K-cluster start solution.

Move an Item to a Cluster

In the nonrandom or deterministic condition, we move item i to the cluster k that yields the greatest improvement of the objective function. Methods that have such a deterministic move (i.e., hill-climbing methods) may provide only a local optimum value; therefore, the success of the method depends on the starting point of the algorithm.

For the random or stochastic condition, we use an adapted Metropolis procedure that is frequently used in simulated annealing (e.g., Berthold & Hand, 1999). The probability P_{it}^{k} that item i will be moved to cluster k at iteration t is derived from the resulting change in the value of the objective function, denoted as ΔO_{it}^{k}, where O_{it}^{k} may refer to O_1, O_2, or O_3. More specifically, the

probability that item i is moved to cluster k equals

$$P_{it}^{k} = \frac{[\exp(\Delta O_{it}^{k})]^{t/I}}{\sum_{i=1}^{I}\sum_{k=1}^{K}[\exp(\Delta O_{it}^{k})]^{t/I}}, \tag{10}$$

where the denominator normalizes the probability to lie within the 0-1 range and sums to 1 over k and i. Power (t/I) is added to make improvements of the value of the objective function more likely for higher iteration numbers. Which particular item i is moved to which cluster k depends on a randomly drawn number and the probabilities described above.

Convergence

When moving items into clusters deterministically, the method stops when the objective function can no longer be improved. With a random component, it is less obvious when to stop the clustering process. We need a convergence rule to stop the iteration process. Figure 1 depicts the value of the objective function as iteration number t increases for some binary multidimensional data. One can observe that for low iteration numbers, the objective function may either increase or decrease. For $t \to \infty$, the algorithm becomes similar to a deterministic algorithm in which the best improvement is always selected. In Figure 1, one can observe that the process does not become completely deterministic because the value of the objective function fluctuates between the best and the next best solution. This is because in our implementation we always move one item at every iteration step.

For convergence of the stochastic move algorithm, one has to let the iteration process continue until a more or less stable result is obtained. In Figure 1 this occurs after approximately 6,000 iterations. The rather safe stopping rule we used in our simulation studies is that the same largest value of the objective function should have occurred 100 times.

Simulation Study

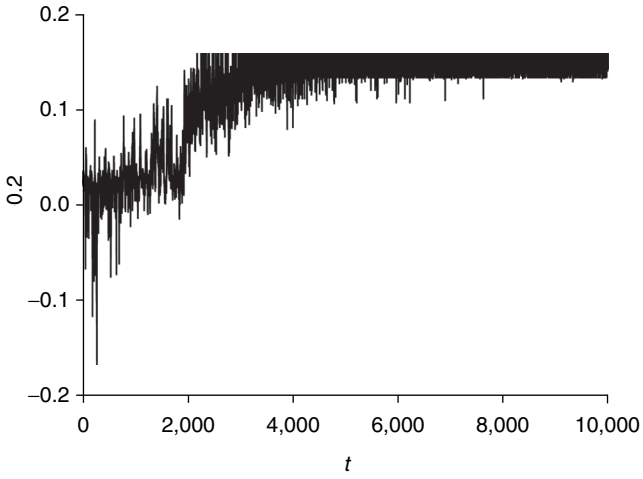
The first goal of the simulation study was to investigate how successful the objective functions O_1, O_2, and O_3 are in assessing the underlying dimensionality structure for different types of data matrices. For this purpose, we compared the solution corresponding to the globally optimal value and the underlying dimensionality of the data. One may recall that the globally optimal value is the highest value obtained after running each algorithm. The second research question was which algorithm can find the global maxima for O_1, O_2, and O_3. In answering this question, we ignored the underlying dimensionality and assessed only which algorithms yielded global solutions. This will be discussed in the second part of this section.

Model Used for Generating Data

For generating binary item scores, we used a model that can produce item response functions (IRFs) with more than one inflection point (also see Douglas & Cohen, 2001; Samejima, 2000). The model is a multidimensional IRT model that consists of a mixture of IRFs (denoted "components") that satisfy the multidimensional two-parameter logistic model (M2-PLM; Birnbaum, 1968; Reckase, 1997).

Before we define the model, we need to introduce some notation. Let q $(q = 1, \ldots, Q)$ represent one of the mixture components; α_{iqd} is the discrimination parameter of component q on trait d $(d = 1, \ldots, D)$ for item i, and δ_{iqd} is the component-specific difficulty parameter for item i. The

Figure 1
Example of Convergence Process When Move Is Random



Note. t denotes iteration number.

component-specific difficulty may be interpreted as the location where according to that component the item concerned discriminates most.

The mixture model is defined by

$$P(X_i = 1|\theta) = \sum_{q=1}^{Q} \frac{\exp[\sum_{d=1}^{D} \alpha_{iqd}(\theta_{pd} - \delta_{iqd})]}{1 + \exp[\sum_{d=1}^{D} \alpha_{iqd}(\theta_{pd} - \delta_{iqd})]}. \tag{11}$$

Increasing the number of components in the model generally means that more inflection points are added to the IRF. Increasing the α_{iqd}s means that the local increases in the IRF become steeper. Because an IRF differs locally in steepness, increasing α_{iqd} may increase the overall discrimination of an IRF. Increasing the α_{iqd}s does not, however, unequivocally manipulate the overall discrimination of an item. The item discrimination can be manipulated more directly via the δ_{iqd}s; that is, increasing the variance of the δ_{iqd}s within an item lowers the discrimination of an item.

In the simulation study, depending on the particular cell in the design, the parameters of the mixture model took on different values. The values of item component parameters were first generated from a distribution and subsequently fixed so that the results of the various conditions of the design became comparable. This means that the values of α_{iqd} and δ_{iqd} are different between items and between components, but the item properties were the same between equivalent conditions of the design. Five components for each IRF were used.

We investigated the relationship between parameters of the five-component mixture model (for three levels of component-sensitive discrimination and two levels of component-sensitive difficulty) and parameters of the M2-PLM (see Table 2). M2-PLM parameters were estimated using the LEM computer software (Vermunt, 1997). For simplicity, unidimensional items were used; all other properties of the items were the same as in the simulation study. Because the values of the parameters were fixed across conditions, this meant that the δ_{iqd}s of items were exactly the same across different levels of component-sensitive discriminations, making it possible to assess the effect of α_{iqd} without the influence of δ_{iqd}. The reverse (i.e., α_{iqd} fixed and δ_{iqd} free) was true for δ_{iqd}.

Table 2
Minimum and Maximum Values of the Two-Parameter Logistic Model Parameters, Item Discrimination, and Item Difficulty for Data Generated Using Two Ranges of Component Difficulty (δ_{iqd}) and Three Levels of Component Discrimination (α_{iqd})

	Level of α_{iqd}		
Range of δ_{iqd}	High	Medium	Low
		Item discrimination	
Small	1.36–4.10	1.27–2.28	0.49–0.89
Large	0.70–1.54	0.64–1.58	0.44–0.87
		Item difficulty	
Small	−1.26–1.32	−1.32–1.22	−1.24–1.35
Large	−1.32–1.68	−1.23–1.36	−0.98–1.41

Design

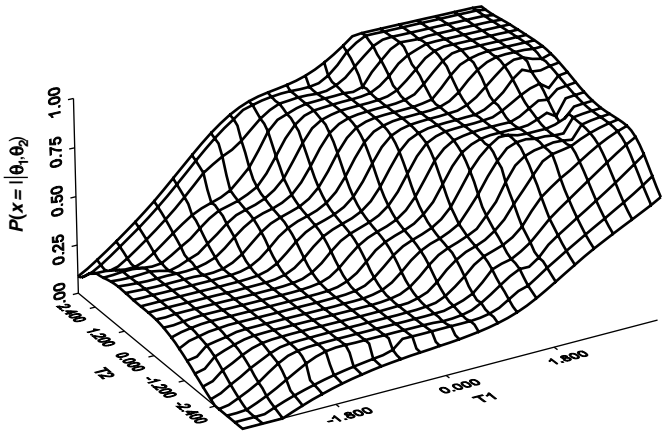
Retrieving the dominant underlying dimensionality. To answer the research question regarding the efficacy of the objective functions, we used data matrices having 15 items per latent trait and 2,000 responses per item. Throughout the study, we used two-dimensional, standard normally distributed latent traits (θ_1 and θ_2). In the main study, five components (in Equation 11) were used. To ensure stability of the results, we replicated each cell 10 times. The design comprised the completely crossed factors Correlation Between Traits (three levels), Structure (three levels), and Item Discrimination (two levels), which yielded a $3 \times 3 \times 2$ design. For part of the design, we investigated the effect of other factors, such as Numbers of Items Per Trait (two levels) and Sample Size (three levels). In addition, for part of the design, a simpler model with one component was used, which allows independent investigation of the effect of the factors Structure and Item Discrimination.

The three levels of Correlation Between Traits (ρ) were .1, .4, and .7. As ρ increases, the less the responses to items sensitive to different traits can be expected to be different, and the more difficult it becomes to find the correct partitioning. The extremes $\rho = .0$ (i.e., no correlation) and $\rho = 1.0$ (i.e., a unidimensional model fits the data best) were not included because they provide little challenge for the methods. van Abswoude et al. (2004) showed that for high-discrimination items, MSP could find the correct dimensionality for $\rho \leq .4$. We would like to know whether better results can be obtained using the new MSA methods.

The three levels of Structure were conditions AS1, AS2, and AS3 (AS stands for approximate simple structure, as used by Stout, 2002). In Condition AS1, items were constructed to be highly discriminating on a single latent trait (i.e., the intended trait). Discrimination with respect to the other latent trait (i.e., the unintended latent trait) was entirely due to the correlation between the traits. In the AS2 condition, items discriminated highly with respect to their intended trait and weakly with respect to their unintended trait. Figure 2 depicts an item response surface used in the AS2 condition. In Condition AS3, items discriminated highly with respect to their intended trait and at a medium level with respect to their unintended trait. We expected that the simulated dimensionality structure in AS1 would be easiest to recover, followed by AS2, and then by AS3. The three levels of Structure were obtained via the specification of the component-specific discrimination parameters.

The two levels of Item Discrimination were high and moderate. The overall discrimination of an IRF was manipulated via the dispersion of the δ_{iqd}s, as was explained earlier. The factor Item

Figure 2
Item Response Surface of an Item That Discriminates Highly With Respect to θ_1 (T1) and Lowly With Respect to θ_2 (T2)



Discrimination was manipulated independently from the factor Structure. Thus, within one level of Structure (say, AS1) the item discrimination may be high or moderate.

The two levels of Numbers of Items Per Trait were equal numbers (i.e., 15 items sensitive to θ_1 and 15 items to θ_2) and unequal numbers (i.e., 15 items sensitive to θ_1 and 30 items to θ_2). Numbers of Items Per Trait was included as a design factor because it has been shown to have an effect on finding the dimensionality of a set of items using LI-based methods (van Abswoude et al., 2004). The effect of this factor was investigated for a few cells.

The three levels of Sample Size we used were small (i.e., 200 subjects), medium (i.e., 2,000 subjects), and large (i.e., 10,000 subjects). It was expected that the results using small sample sizes would be less stable than the results for medium or large sample sizes. The effect of this factor was investigated for only a few cells.

Dependent variables. Judgment about the success of the methods was based on two criteria. The first criterion was whether the simulated structure was recovered, meaning that the items were split according to their underlying trait structure. The retrieved clustering solution has the following general structure $[K : I^1; I^2; \ldots I^K]$, where K denotes the number of obtained clusters and $I^1, \ldots, I^K$ denotes the number of items retrieved for clusters 1 through K. When no classification errors were made, the two adjacent clusters (e.g., I^1 and I^2) are separated by a semicolon; separation by a comma is used to indicate that the two adjacent clusters are in fact sensitive to the same underlying trait, and separation by a slash indicates that some items are entered into a cluster sensitive to another trait. When a method yields the simulated structure, this is referred to as the "true dimensionality," which is an observation rather than an interpretation. For example, in the case of two latent variables with a correlation of .95, from a substantive point of view one may prefer the incorrect solution [1:30] over the correct partitioning [2:15;15].

The second dependent variable was the value of the objective function at the global maximum solution. Note that the value of the objective function provides us with an indication of the strength of the clusters that were found.

Performance of the algorithms. The second research question relates to the elements of the clustering method that are responsible for finding global optima. For this purpose, we compared the various variants of the NHCA method with respect to their ability to find the global maxima for different types of data. We used only a part of the total design. We investigated the effect of Structure (four levels), and Correlations Between Latent Traits (three levels) on the number of times the global optimum was found for the different NHCA algorithms. The four algorithms (i.e, two types of initial configurations and two move processes) were investigated for O_1, O_2, and O_3 each. The probabilistic methods were run 10 times. The data were generated using the five-component mixture model with moderately discriminating items.

Dependent variables. The success of the algorithms was judged by using as a criterion the number of replications that resulted in a global optimal solution. We also determined the average number of runs (and its standard deviation) producing the global optimum solution for 10 replicated data matrixes. The average number of iterations needed to find the global optima was the second dependent variable.

Results

Retrieving the Dominant Underlying Dimensionality

Table 3 shows the dimensionality results using sequential clustering (in short, Sequent) and NHCA for data ($N = 2{,}000$) generated with the five-component mixture model (narrow distributions of the component difficulties). For Sequent, we present the clustering solution. For nonhierarchical clustering, we present the value of the objective function at the global maximum and at the true dimensionality (within parenthesis). In Table 3, the label ''true'' denotes that the simulated dimensionality was retrieved (i.e., the global solution is the same as the true dimensionality); ''otherwise, the retrieved global clustering solution is printed.

As Correlations Between Latent Traits (ρ) increased, the following effects can be observed in Table 3: Sequent tended to collect all items into one large cluster; O_1 and O_2 found the underlying dimensionality; and O_3 tended to split the total set in one item pair versus the rest. When loadings on unintended traits increased (from AS1 to AS3), the intended dimensionality was obtained less often. For moderate discrimination, Sequent and to a lesser extent O_1 did not yield the true dimensionality, whereas O_2 did in most cases. The value of the objective function increased with increasing ρ, increasing loading on unintended traits, and increasing Item Discrimination. In general, O_1 and O_2 performed better than Sequent.

In Sequent, for high ρ and deviation from AS1, most items satisfied the Mokken scale conditions for the first scale, which means that they were all collected into the first scale (see Table 3). This effect was stronger in the high-discrimination condition. With high ρ and moderate discrimination, some items could not satisfy the scale conditions of the large cluster, but satisfied the scale conditions when a new scale was formed (an example can be found in a later table) or were not scalable at all. Item and scale H values were lower with the moderate discrimination because the H coefficient is sensitive to item discrimination.

The difference between the global maximum value of the objective function and its value for the true dimensionality partitioning provides some indication as to the suitability of the method for dimensionality analysis. One may note that the global maximum value is at least as large as the value at the true dimensionality. Equal or similar values indicate that the clustering method is suitable for dimensionality assessment, which is exactly what we observe for O_1 and O_2, with exception of some combinations of AS3, high correlation between traits and moderate discrimination. The values for

Table 3
Results of the Restricted Sequential MSP and the Unrestricted New Mokken Scale Analysis Methods Using Objective Functions O_1, O_2, and O_3 on Retrieved Dimensionality

			NHCA					
		Sequent ($c = .3$)	$O_1(H_{ij} - \bar{H}_{ij})$		$O_2(\bar{H}_{ij})$		$O_3(H^k)$	
Structure	ρ	Clust.	Value	Clust.	Value	Clust.	Value	Clust.
				High item discrimination				
AS1	.1	True	.290	True	.573	True	.543(.561)	[2/28]
	.4	[14/16]	.193	True	.560	True	.680(.548)	[2/28]
	.7	[30]	.089	True	.563	True	.728(.552)	[2/28]
AS2	.1	True	.210	True	.529	True	.658(.514)	[2/28]
	.4	[29]	.127	True	.543	True	.699(.528)	[2/28]
	.7	[30]	.067	True	.562	True	.738(.551)	[2/28]
AS3	.1	[28]	.070(.065)	[2/25]	.471	True	.692(.451)	[2/28]
	.4	[30]	.068(.039)	[6/24]	.500	True	.717(.482)	[2/28]
	.7	[30]	.067(.023)	[6/24]	.542(.540)	[15/15]	.746(.524)	[2/28]
				Moderate item discrimination				
AS1	.1	[2,9;10]	.129	True	.256	True	.383(.256)	[2/28]
	.4	[9;9]	.078	True	.246	True	.417(.245)	[2/28]
	.7	[9;2,9]	.050(.041)	[7/23]	.260	True	.426(.258)	[2/28]
AS2	.1	[2,6;2,5]	.080	True	.217	True	.377(.216)	[2/28]
	.4	[2,7;9]	.054	True	.226	True	.383(.226)	[2/28]
	.7	[8;2,8]	.044(.027)	[8/22]	.241	True	.412(.241)	[2/28]
AS3	.1	[2,4;2,5]	.050(.018)	[7/23]	.179(.177)	[12/18]	.364(.178)	[2/28]
	.4	[2/2/2/9]	.051(.014)	[7/23]	.216(.211)	[13/17]	.388(.211)	[2/28]
	.7	[2/2/2/12]	.053(.007)	[6/24]	.243(.229)	[13/17]	.443(.229)	[2/28]

Note. NHCA = nonhierarchical clustering analysis; Clust. = the clustering solution of a method; value = the near global objective functions' value and, if different, the value at the true dimensionality (within parenthesis); true = the maximum value of the objective function was found at the simulated partitioning (otherwise, the obtained partitioning is presented in notation between brackets); semicolons separate dimensionally different sets of items; commas separate dimensionally similar sets; and slashes separate dimensionally mixed sets.

O_3 were so far apart that we can safely conclude that O_3 was not successful in finding the underlying dimensionality.

One may note that low O_1 values were found when items discriminated weakly, clusters correlated highly, or items loaded highly on each latent trait (see Table 3). In those cases, the pairwise H_{ij}s do not deviate much from their mean value. When we use Mokken's rules to interpret obtained O_2 values (also see Discussion), we observe that the clusters obtained were strong for high-discriminating items and very weak for moderately discriminating items (see Table 3). Interpreting the maximum O_3 values using Mokken's rules of thumb indicates that the average cluster is strong for high-discriminating items and weak to medium for moderately discriminating items.

As ρ increased and as the item loadings on unintended traits increased, it became more difficult to retrieve the simulated dimensionality structure. Whether this can be attributed to sampling fluctuation is investigated next.

Table 4
Results of the Unrestricted New Mokken Scale Analysis Methods Using Objective Functions O_1, O_2, and O_3 Concerning Retrieved Dimensionality for 10 Replicated Data Matrices

		NHCA					
		$O_1(H_{ij} - \bar{H}_{ij})$		$O_2(H_i^k)$		$O_3(H^k)$	
Structure	ρ	*M*(*SD*)	# True	*M*(*SD*)	# True	*M*(*SD*)	# True
				High item discrimination			
AS1	.1	.288(.008)	10	.556(.009)	10	.403(.018)	0
	.4	.192(.007)	10	.557(.006)	10	.392(.037)	0
	.7	.096(.006)	10	.559(.009)	10	.427(.021)	0
AS2	.1	.203(.005)	10	.523(.008)	10	.360(.018)	0
	.4	.128(.007)	10	.538(.008)	10	.401(.016)	0
	.7	.064(.003)	10	.555(.007)	10	.429(.016)	0
AS3	.1	.072(.004)	0	.461(.013)	9	.340(.016)	0
	.4	.070(.003)	0	.500(.011)	10	.388(.025)	0
	.7	.065(.003)	0	.543(.008)	4	.431(.013)	0
				Moderate item discrimination			
AS1	.1	.125(.005)	10	.257(.004)	10	.403(.018)	0
	.4	.085(.004)	10	.253(.004)	10	.371(.044)	0
	.7	.047(.002)	1	.253(.005)	10	.383(.047)	0
AS2	.1	.078(.004)	10	.220(.006)	10	.315(.036)	0
	.4	.054(.003)	10	.230(.005)	10	.364(.042)	0
	.7	.050(.003)	0	.247(.005)	8	.386(.042)	0
AS3	.1	.052(.003)	0	.177(.006)	0	.297(.040)	0
	.4	.055(.003)	0	.211(.007)	0	.340(.034)	0
	.7	.059(.005)	0	.244(.004)	0	.414(.030)	0

Note. NHCA = nonhierarchical clustering analysis; *M*(*SD*) = average and standard deviation of the maximum objective function for 10 replicated data matrices; # true = number of correct partitions for the 10 replicated data matrices.

Replicated data. To get some idea about the stability of the results, the analyses were repeated with 10 different samples ($N = 2{,}000$). Table 4 shows that the results were not sensitive to sampling fluctuations. The maximum values of the objective function did not change much between replications. In the more difficult AS3 and $\rho = .7$ conditions, O_1 and O_2 were maximized at the true partitioning in some but not all replications.

Number of items and sample size. Table 5 shows the results obtained when varying the factors Number of Items and Sample Size for moderately discriminating items. These data were simulated under the AS2 condition. For Sequent, O_1, O_2, and O_3, the results of 10 replicated data matrices are presented.

Table 5 shows that the numbers of items did not influence whether Sequent found the true dimensionality. For unequal numbers of items per trait, Sequent generally produced more clusters than for equal numbers of items (not shown in Table 5). The results of O_1 were worse in the unequal numbers of items condition than in the equal numbers of items condition, but they turned

Table 5
Results of the Unrestricted New Mokken Scale Analysis Methods Using Objective Functions O_1, O_2, and O_3 Concerning Dimensionality Results for Different Numbers of Items Per Trait and Number of Respondents for Moderate Discrimination Data

		NHCA					
		$O_1(H_{ij} - \bar{H}_{ij})$		$O_2(H_i^k)$		$O_3(H)$	
	ρ	*M*(*SD*)	# True	*M*(*SD*)	# True	*M*(*SD*)	# True
Default							
	.1	.078(.004)	10	.220(.006)	10	.315(.036)	0
	.4	.054(.003)	10	.230(.005)	10	.364(.042)	0
	.7	.050(.003)	0	.247(.005)	8	.386(.042)	0
Numbers of items per trait							
[2:15;30]	.1	.077(.005)	10	.219(.009)	10	.379(.020)	0
	.4	.052(.010)	1	.228(.004)	10	.409(.018)	0
	.7	.030(.013)	0	.243(.006)	7	.440(.022)	0
Sample size							
200	.1	.075(.007)	2	.218(.021)	4	.463(.038)	0
	.4	.058(.008)	1	.227(.024)	3	.466(.037)	0
	.7	.060(.006)	0	.250(.016)	0	.533(.045)	0
10,000	.1	.079(.002)	10	.221(.003)	10	.358(.010)	0
	.4	.053(.002)	10	.232(.003)	10	.384(.013)	0
	.7	.050(.001)	0	.241(.004)	10	.414(.009)	0

Note. NHCA = nonhierarchical clustering analysis; *M*(*SD*) = average and standard deviation of the maximum objective function for 10 replicated data matrices; # true = number of correct partitions for the 10 replicated data matrices.

out to be better than expected. Evidently, the effect of having clusters that are not equal in strength (i.e., due to unequal numbers of items or unequal item discrimination between clusters) is not so large for two-cluster data. The results for O_2 and O_3 were not notably different for the equal and unequal numbers of items conditions.

There is some sample fluctuation in Sequent's results, but not to the extent that simulated partitionings were retrieved in one condition and not in the other. Table 5 shows that with samples of size 200 the values of the objective functions varied more across replications and the true dimensionality was found less often than with the larger sample sizes, which is, of course, what could be expected.

Additional M2-PLM simulations. To ensure that the obtained results are not an artifact of the mixture model, the O_1, O_2, and O_3 methods were exposed to M2-PLM data. Equation 11 with $Q = 1$ equals the M2-PLM. With this simpler model, it is easier to control factors such as Discrimination and Structure than with the mixture model. This allows us to vary only the intended factor while keeping the remaining factors constant. Unbalanced data, in which the number of items, item discrimination, and/or the distribution of the item difficulties is not the same between clusters, were also developed with the M2-PLM. Unbalanced data sets were generated because they can provide more challenge for the methods than balanced data sets (see Objective Functions Using the *H*

Coefficient). By default, the M2-PLM data has the following properties: $\rho = .1$ (Correlation Between Traits), AS2 (Structure), high (Discrimination), large (Distribution of Difficulty), [2:15;15] (Numbers of Items Per Trait), and $N = 2{,}000$ (Sample Size). We did not replicate.

To demonstrate that our methods capture real dimensionality and not just sampling fluctuation, we also performed an exploratory factor analysis (EFA) for categorical variables (Mplus; Muthén & Muthén, 2003). We used a means and variance-adjusted weighted least squares (WLSMV) method, which provides a χ^2 goodness-of-fit test. In the EFA, the number of factors was fixed to two, and the factors were obliquely (Promax) rotated to allow for the correlation between the latent variables ($\rho \geq .1$). For comparison with our methods, an item is assigned to the factor for which it has the highest loading.

Results. Table 6 presents the M2-PLM results for O_1, O_2, O_3, and EFA. For EFA, the obtained estimated correlations between the two factors, the obtained partitioning according to the two-factor solution, three fit measures—$\chi^2 p$ values RMSR $< .05$ (†) and $> .05$ or RMSEA $> .06$ (‡)—are given.

In general, O_1 and O_2 yielded the true dimensionality more often for the M2-PLM data than for the mixture model data (see Table 6). In particular, for moderate discrimination, AS3, and $\rho = .7$, the obtained differences are noteworthy. The different shape of the item response surfaces for the M2-PLM and the mixture model may explain the obtained results. Alternatively, the M2-PLM data being cleaner than the mixture model data could have had an effect. In particular, using the mixture model item discrimination with respect to each trait is difficult to control. For the mixture model, properties this property is ignorable because it yields acceptable multidimensional nonparametric IRT surfaces. However, for AS3 and $\rho = .7$, this can yield monotonicity violations (for an example, see Figure 2). These properties could hamper finding the preferred solution. With the M2-PLM data, where there were no monotonicity violations, the preferred dimensionality was found. Neither distribution of difficulty nor balanced versus unbalanced item pools had noticeable effects on the performance of the methods.

The effect of factors of the main design on O_1, O_2, and O_3 values was similar for data generated under the M2-PLM and the mixture model. Thus, the M2-PLM results do not give cause to readjust the conclusions about O_1, O_2, and O_3 values. The unbalanced versus balanced conditions did not affect O_1, O_2, and O_3 values. As expected, the O_1, O_2, and O_3 values increased as the Distribution of Difficulty increased.

EFA generally yielded the same dimensionality solutions as O_1 and O_2 (see Table 6). The Promax rotated two-factor model did not fit well for $\rho = .7$, AS3, $N = 200$, and unbalanced data. These EFA results generally confirm the new MSA method's replication results, as discussed earlier (see Table 4). The effect of sample size on factor analysis solutions has been well documented (e.g., Hoogland, 1999). One may note that the difference between ρ and the estimated correlation (Corr) can be explained by deviation from simple structure in the data. For example, the default data set that has $\rho = .1$ and AS2 yielded Corr $= .253$, whereas another data set with $\rho = .1$ and AS1 yielded Corr $= .126$. Thus, the correlations obtained with EFA reflect the simulated structure well (Table 4).

Performance of the Algorithms

The results of the new methods presented in Tables 3 through 6 were obtained by making use of all NHCA algorithms. Only the results that corresponded with maximum values of the objective functions were presented. In this section, the goal is to find out which algorithm was best at finding these approximately global solutions. The algorithm that had the highest probability of finding the

Table 6
Results of O_1, O_2, O_3, and Exploratory Factor Analysis for Balanced and Unbalanced Clusters Simulated Using the Multidimensional Two-Parameter Logistic Model

	$O_1(H_{ij} - \bar{H}_{ij})$		$O_2(H_i^{(k)})$		$O_3(H^{(k)})$		EFA	
Correlation Between Traits (ρ)	Value	Clust.	Value	Clust.	Value	Clust.	Value	Clust.
ρ = .1 (default)	.217	True	.531	True	.617(.512)	[2:2/28]	.253	True
ρ = .4	.137	True	.564	True	.701(.546)	[2:2/28]	.509	True
ρ = .7	.064	True	.575	True	.741(.557)	[2:2/28]	.690	True†
ρ = .1/.4	.179	True	.566	True	.686(.548)	[2:2/28]	.403	True
Structure								
AS1	.260	True	.529	True	.541(.512)	[2:2/28]	.126	True
AS2 (default)	.217	True	.531	True	.617(.512)	[2:2/28]	.253	True
AS3	.063	True	.613	True	.766(.601)	[2:2/28]	.666	[2:14/16]†‡
[2:hi/lo;me/lo]	.142	True	.384	True	.611(.373)	[2:2/28]	.349	True†
Discrimination								
Moderate	.066	True	.220	True	.316(.217)	[2:2/28]	.464	True
High (default)	.217	True	.531	True	.617(.512)	[2:2/28]	.253	True
Distribution of difficulty								
Small	.184	True	.479	True	.518(.476)	[2:2/28]	.268	True
Large (default)	.217	True	.531	True	.617(.512)	[2:2/28]	.253	True
Combined	.187	True	.498	True	.533(.496)	[2:2/28]	.244	True†
Sample size								
200	.194	True	.550	True	.668(.527)	[2:2/28]	.327	True†‡
2,000 (default)	.217	True	.531	True	.617(.512)	[2:2/28]	.253	True
#Items per trait								
[2:15;15] (default)	.217	True	.531	True	.617(.512)	[2:2/28]	.253	True
[2:15;30]	.210	True	.526	True	.678(.526)	[2:2/43]	.295	True†

Note. EFA = exploratory factor analysis; value = objective functions' (near) global value (and, if different, the value at the true dimensionality in parenthesis); clust. = clustering solution of a method; true = maximum value of the objective function was found at the simulated partitioning (otherwise, the obtained partitioning is presented in notation between brackets); / = dimensionally mixed sets; † = p value < .05; ‡ = RMSEA > .06 or RMSR > .05. Unbalanced clusters appear in bold.

global solution is seen as the best algorithm. When several algorithms performed equally well, we prefer the one that finds the global optimum within the smallest number of iterations.

Table 7 presents the performance of the algorithms used for O_1 and O_2. We did not include O_3 because it was not very successful in unrestricted dimensionality assessment. The algorithms are abbreviated in the following way: SEQ and RAN1 denote the sequential and random initial configurations, and DET and RAN2 denote the deterministic and random moves. The four combinations of the two initial configuration possibilities and the two move possibilities yield the four studied

Table 7
Efficiency of Algorithms for O_1 and O_2

		Initial: Random (RAN1)						Sequential (SEQ)				
		Move: DET			RAN2			DET		RAN2		
Structure	ρ	G	$M(SD)$	$\bar{t}$	G	$M(SD)$	$\bar{t}$	G	$\bar{t}$	G	$M(SD)$	$\bar{t}$
AS1	.1	10	9.1(1.0)	14	10	10.0(0.0)	2166	10	7	10	9.2(0.9)	2215
	.4	10	9.4(1.0)	14	10	9.7(0.7)	3143	10	7	10	9.7(0.7)	3116
	.7	10	7.3(2.1)	13	10	9.3(1.3)	5698	1	8	10	9.5(0.8)	5683
AS2	.1	10	9.2(1.1)	14	10	9.5(0.5)	3453	10	9	10	9.3(0.7)	3385
	.4	10	5.0(1.7)	13	10	10.0(0.0)	5026	10	9	10	9.9(0.3)	5143
	.7	10	8.8(1.7)	14	10	8.5(1.5)	5357	4	11	10	8.3(1.8)	5361
AS3	.1	10	9.6(1.0)	14	10	9.1(0.7)	5217	9	11	10	9.4(0.7)	5104
	.4	10	9.4(1.1)	14	10	8.2(1.5)	4793	9	16	10	8.7(0.7)	4715
	.7	10	9.6(1.3)	14	10	7.6(1.7)	4562	10	11	10	7.8(1.9)	4675
AS1	.1	10	10.0(0.0)	14	10	9.0(1.5)	2264	10	7	10	8.9(1.0)	2219
	.4	10	10.0(0.0)	14	10	9.3(1.3)	3208	10	7	10	9.3(0.8)	3184
	.7	10	10.0(0.0)	13	10	9.9(0.3)	6040	10	8	10	9.5(0.5)	6148
AS2	.1	10	10.0(0.0)	14	10	8.7(1.2)	3457	10	4	10	9.1(1.1)	3427
	.4	10	10.0(0.0)	14	10	9.4(0.7)	4932	10	6	10	9.2(1.2)	4826
	.7	10	8.7(2.2)	14	10	7.7(2.5)	8491	6	6	10	7.8(1.8)	8618
AS3	.1	5	2.3(3.2)	13	5	1.3(2.2)	8772	4	3	4	2.7(3.1)	6953
	.4	4	2(3.0)	13	7	0.8(0.6)	8317	1	5	7	0.8(0.6)	7149
	.7	3	4(4.8)	14	7	1.3(1.2)	8423	2	5	8	1.2(1.1)	7725

Note. G = number of global maxima out of 10 replications; $M(SD)$ = average and standard deviations of reported global maxima; $\bar{t}$ = average number of iterations over 10 runs of the algorithms and 10 replicated data matrices having moderate discrimination items.

algorithms: RAN1&DET, RAN1&RAN2, SEQ&RAN2, and SEQ&DET. For each algorithm, we present the number of times out of 10 replications that the highest (global) maximum was reached (denoted G), the average number of runs that yielded global values over 10 replications and their standard deviations (denoted $M(SD)$, not shown for SEQ & DET), and the average number of iterations needed to obtain the global solution for the first time (denoted t).

Table 7 shows that the algorithms with a random component (i.e., RAN1&DET, RAN1 &RAN2, and SEQ&RAN2) performed best at finding global solutions for O_1 (i.e., the sums of G were 120 for each algorithm). For O_2, RAN1&RAN2 yielded the global solution 105 times, SEQ&RAN2 106 times, and RAN1&DET 101 times. The completely deterministic algorithm performed worst for the two objective functions: It found the global solution 91 times for O_1 and 72 times for O_2. The values presented within parentheses in Table 7 tell us that none of the algorithms yielded the global solutions every time they were run, but the random-move algorithms came closest. One may note that as ρ increased and as items loaded on more than one trait, the number of times a global solution was obtained decreased and the number of iterations needed to obtain a solution increased.

The relationship between the discrete partitionings and the objective function's value can explain the results of Table 7. The table shows that for low ρ and weak loadings on unintended traits (AS1 and AS2), all algorithms yielded the global solution. In these data matrices, the

relationship between the partitionings and the value of the objective function (for O_1 and O_2) is relatively simple because there is only one maximum, and the fast deterministic algorithm can be used to find global maxima. For data having high ρ and high loadings on unintended traits (AS3), the relationship between partitionings and the objective function is more complex because in addition to the global maximum, there are local maxima, and the values of these maxima may be close to one another (see Table 3). The complexity also means that the partitionings yielding these maxima may not at all be similar; that is, many items need to be moved before a different maximum is found. For this type of data, stochastic algorithms are needed.

The choice between the random-start/deterministic move and the random-move algorithms is a matter of taste. The RAN1&DET algorithm, however, won this match. It is an attractive algorithm because of its high accuracy and high efficiency. The random-move algorithms, which search a large part of the solution space, were highly accurate but required many iterations to converge.

Conclusions

Three new MSA approaches to resolve the optimization problems of sequential MSP were introduced in this article. The objective functions in these new methods incorporate reformulations of multidimensional Mokken scaling, and deterministic and stochastic nonhierarchical clustering algorithms were used to maximize these objective functions. To investigate the properties of the objective functions, we ignored the restrictions that are usually imposed in a Mokken scaling analysis.

The first research question that we wanted to answer was "How successful are the three new objective functions at finding the underlying dimensionality of a data set?" Objective function O_2 yielded the best results; O_1 performed somewhat better than the original sequential approach, and O_3 performed worse than the original sequential approach. Also, the fact that the new methods using O_1 and O_2 were able to find the true dimensionality in most situations indicates that these are effective tools for dimensionality assessment, perhaps even comparable to methods based on weak LI (e.g., Stout, 2002). This confirms earlier findings that the H coefficient can be used not only as a tool for scaling but also for dimensionality assessment (Hemker et al., 1995; van Abswoude et al., 2004).

The methods using O_1 and O_2 performed approximately equally well in most conditions of the study. This is not surprising because the two are strongly related. However, there are some differences. Objective function O_1 has the advantage that under $D = 1$ (i.e., unidimensionality) $O_1 = 0$; therefore, deviation from unidimensionality can be determined. Another advantage of O_1 is that theoretically its value is maximized when the investigated number of traits (K) equals the true underlying number of traits (D). Objective function O_2 does not have this advantage, but this can easily be remedied. A disadvantage of O_1 is that this objective function may not work well when D is large and when clusters have unequal numbers of items or have differently discriminating items. The impact of these disadvantages requires further investigation. The interpretation of O_2 as the average within-cluster H_i is simpler than the interpretation of O_1. Therefore, it is likely that known properties of MSA (e.g., Mokken's rules of thumb) can be used. Both O_1 and O_2 use the H coefficient; therefore, both join items on the basis of the slope of the IRFs. Based on the results from the simulation study and the properties discussed above, we prefer O_2 to O_1.

The second research question was "Which algorithm should we use in the new MSA?" A completely deterministic algorithm should clearly not be used because it can find only global maxima for simple structure data with slightly correlated or uncorrelated latent traits. For the preferred objective function O_2, the algorithm with the random-start configuration and the deterministic move performed well, that is, with high precision and high speed.

Discussion

Some issues deserve further attention. First, providing mechanisms for deciding between several globally optimal solutions when the true number of latent traits is unknown was not the aim of this study. However, one could easily imagine that for this purpose different sources of information could be used. In particular, the value of the objective function (see second issue), the item and scale H values, the purpose of scaling, and substantive information (see third issue) are informative. For explicit rules of thumb (say, cutoff scores for the objective functions that hold under various conditions), an additional simulation study would be required that explicitly targets this research question. Such a study most probably would be restricted to one objective function (say, O_2) and one algorithm (say, RAN1&DET).

Second, the following preliminary rules of thumb for interpreting the values of O_1 and O_2 are based on the M2-PLM and mixture model data. For O_1, we propose that when $O_1 \leq .05$, interpret with caution; when $.05 \leq O_1 \leq .1$, there is adequate scaling; and when $.1 \leq O_1 \leq 1$, there is very promising scaling. The motivation behind the word of caution is that theoretically $O_1 = 0$ when $D = 1$. In addition, low values were found especially for AS3, $\rho = .7$, and moderate discrimination and can therefore be indicative of chance capitalization (the assignment of items to clusters on the basis of noise). Joining clusters may be a strategy that yields better results (i.e., higher O_1 values) when $O_1 \leq .5$. For O_2, we propose to use Mokken's rules of thumb, which were presented earlier. Although there are clear distinctions between H and O_2, O_2 can be seen as an average unweighed H over K clusters (see Equation 8). Thus, there are enough similarities that make the application of the rules defensible.

Third, when confronted with two highly correlated sets or sets with high loadings on unintended traits, researchers may differ in their opinions as to whether these sets should be joined. The NHCA methods can be applied whether a researcher prefers to join items or not. The decision about the number of clusters to use is left to the researcher.

Fourth, in MSA with the MSP software (Molenaar & Sijtsma, 2000), items will automatically satisfy Mokken scale conditions. We left the Mokken scale conditions out of the NHCA methods because we wanted to know whether the methods could be used to find globally optimal solutions and whether these solutions reflected the simulated dimensionality structure. If the Mokken scale conditions (i.e., with $c = .3$) were incorporated, weakly scalable items, for which the assignment of items into clusters is the most difficult, would have been left out of the analysis; thus, the limitations of the methods would have been difficult to investigate. Future research will address how the Mokken scale conditions can be incorporated into the new MSA methods.

Fifth, variations of the objective functions can be obtained by changing the definition of η. For example, in Equation 8 we used the average within-cluster H_i. One possible alternative objective function would maximize the average within H_i and minimize the H_i between clusters simultaneously, which could be achieved simply by equating η_i^k to -1 rather than to 0 when item i is not in cluster k. Although this objective function no longer has the convenient interpretation of the average within-cluster H_i, it may be an appropriate objective function for determining the number of clusters. This is because the objective function contains a punishment when items that should be in the same cluster are put in separate clusters. These issues will be addressed in future research.

Note

1. Number of items, underlying dimensions, and scale H are presented for each scale.

References

Berthold, M., & Hand, D. J. (1999). *Intelligent data analysis: An introduction*. New York: Springer.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Danzig, G. B., & Thapa, M. N. (1997). *Linear programming*. New York/Berlin: Springer-Verlag.

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*, 234-243.

Emons, W. H. M. (2003). *Detection and diagnosis of misfitting item-score patterns*. Amsterdam, Netherlands: Dutch University Press.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383-392.

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

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337-352.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331-347.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, *14*, 1523-1543.

Hoogland, J. J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen, Netherlands.

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

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277-298.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

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

McDonald, R. P. (1985). *Factor analysis and related models*. Hillsdale, NJ: Lawrence Erlbaum.

Meijer, R. R. (1994). *Nonparametric person fit analysis*. Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam, Netherlands.

Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs*. New York: Springer.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: Mouton.

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

Molenaar, I. W., & Sijtsma, K. (2000). Users manual MSP5 for Windows. A program for Mokken scale analysis for polytomous items [Computer software manual]. Groningen, Netherlands: iec ProGAMMA.

Muthén, L., & Muthén, B. (2003). Mplus version 2.13 [Computer software]. Los Angeles: Authors.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41-68.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

Roskam, E. E., Van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: a critical discussion. *Applied Psychological Measurement*, *10*, 265-277.

Roussos, L., & Ozbek, O. (2003, April). *Formulation of the DETECT population parameter and evaluation of DETECT estimator bias*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical

cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1-30.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetrical item characteristic curves. *Psychometrika*, *65*, 319-335.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: Sage.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Stout, W. F. (2002). Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika*, *67*, 485-518.

Stout, W., Goodwin Froelich, A., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). New York: Springer.

van Abswoude, A. A. H. (2004). *Dimensionality assessment under nonparametric IRT models*. Unpublished doctoral dissertation, Tilburg University, Netherlands.

van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*, 3-24.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data [Computer software]. Tilburg, Netherlands: Tilburg University.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213-249.

Zhang, Y. O., Yu, F., & Nandakumar, R. (2003, April). *The impact of conditional scores on the performance of DETECT*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Acknowledgments

The first author performed this research as part of her PhD dissertation project at Tilburg University in the Netherlands. The authors would like to thank Klaas Sijtsma for his useful comments on earlier drafts of this article.

Author's Address

Address correspondence to Alexandra A. H. van Abswoude, Department of Methodology and Statistics, Maastricht University, P.O. Box 616, 6200MD Maastricht, Netherlands; e-mail: a.vanabswoude@stat.unimaas.nl.