Mokken Scale Analysis Using Hierarchical Clustering Procedures

Alexandra A. H. van Abswoude, Tilburg University
Jeroen K. Vermunt, Tilburg University
Bas T. Hemker, Citogroep
L. Andries van der Ark, Tilburg University

Mokken scale analysis (MSA) can be used to assess and build unidimensional scales from an item pool that is sensitive to multiple dimensions. These scales satisfy a set of scaling conditions, one of which follows from the model of monotone homogeneity. An important drawback of the MSA program is that the sequential item selection and scale construction procedure may not find the dominant underlying dimensionality of the responses to a set of items. The authors investigated alternative hierarchical item selection procedures and compared the performance of four hierarchical methods and the sequential clustering method in the MSA context. The results showed that hierarchical clustering methods can improve the search process of the dominant dimensionality of a data matrix. In particular, the complete linkage and scale linkage methods were promising in finding the dimensionality of the item response data from a set of items. *Index terms: item selection methods, multidimensionality, Mokken scale analysis, hierarchical cluster analysis, nonparametric item response theory.*

In the past decade, there has been an increasing interest in using nonparametric item response theory (NIRT) as a tool in test dimensionality assessment. For example, Douglas, Kim, Roussos, Stout, and Zhang (1999) investigated the dimensionality of the Law School Admission Test, and Scheirs and Sijtsma (2001) investigated the dimensionality of data from the International Survey of Adult Crying. In NIRT, it is assumed that all items are sensitive to a single latent trait (unidimensionality, UD), that item responses given this latent trait are statistically independent (local independence, LI), and, for dichotomously scored items, that the probability of answering an item correctly is a monotone nondecreasing function of the latent trait (monotonicity, M) (Mokken, 1971; Sijtsma, 1998). A set of item response data that are UD, M, and LI is denoted as being monotonely homogeneous (MH; Mokken, 1971).

Methods that are aimed at selecting sets of items that are sensitive to one latent trait from larger item pools sensitive to multiple latent traits frequently use relaxations of the UD, M, or LI assumptions and often focus on one or more of these weakened assumptions (e.g., Ip, 2001; van Abswoude, van der Ark, & Sijtsma, 2004). HCA/CCPROX (e.g., Roussos, Stout, & Marden, 1998), DIMTEST (e.g., Nandakumar & Stout, 1993), and DETECT (e.g., Zhang & Stout, 1999a, 1999b) concentrate on the LI assumption. The program discussed in this article, MSP (Molenaar & Sijtsma, 2000), concentrates on the M assumption. The MSP procedure for dichotomously scored items is now discussed.

Applied Psychological Measurement, Vol. 28 No. 5, September 2004, 332–354
DOI: 10.1177/0146621604265510

The Mokken scale analysis method (MSA; e.g., Mokken, 1971; Molenaar & Sijtsma, 2000) and its program (MSP) have, apart from being generally available for applied researchers, a number of attractive properties: If the MH model is satisfied, simple sum scores may be used to stochastically order subjects on the underlying variable (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997); it allows the user to choose only items with sufficient discrimination power into a test (Sijtsma, 1998); it is suitable in contexts where there are, compared to weak-LI methods, few items and few subjects; and the item selection procedure runs quite fast. MSP contains a sequential item selection method, which can be seen as a sequential clustering algorithm. This clustering method has, however, an important drawback: It does not always find back the correct dimensionality structure of the data in multitrait situations (van Abswoude et al., 2004).

In this study, it is investigated whether hierarchical clustering (e.g., Everitt, Landau, & Leese, 2001) algorithms can find the correct dimensionality structure better than the sequential clustering algorithm—that is, whether suboptimal solutions can be prevented in multitrait situations. Four types of hierarchical clustering methods were used to create nonoverlapping sets of dichotomous items, each of which satisfies certain scaling conditions. In a simulation study and an empirical example, these four methods were compared with each other and with MSP's sequential procedure in their ability to find the correct dominant underlying dimensionality structure in situations with more than one latent trait.

Mokken Scale Analysis

Mokken scale analysis is a method that may be used to select a subset of items sensitive to the same underlying dimension from a larger item pool. Similarly to factor analysis, MSA starts with a matrix containing information on the strength of the bivariate relationships between the J items under study. In factor analysis, this is either a correlation or covariance matrix. MSA uses a matrix with H-coefficients (Loevinger, 1948; Mokken, 1971). Let F_{jk} represent the observed number of Guttman (1950) errors for item pair (j, k), and let E_{jk} be the expected number of Guttman errors under the null model of marginal independence. If item j is easier than item k (i.e., more correct answers), a Guttman error occurs when the more difficult item k is answered correctly and the easier item j is answered incorrectly. The pairwise H_{jk} for item pair (j, k) is defined as

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

Let X_j be a binary random variable, denoting an individual's score on item j ($j = 1, \ldots, J$). An item score may take on the value 0 or 1, and let N be the sample size. Furthermore, let $\pi_j = P(X_j = 1)$, $1 - \pi_j = P(X_j = 0)$, and $\pi_{jk} = P(X_j = 1, X_k = 1)$. Items are ordered such that $\pi_j > \pi_k$, for all $j > k$. Noting that $F_{jk}/N = \pi_k - \pi_{jk}$ and $E_{jk}/N = (1 - \pi_j)\pi_k$, H_{jk} can also be written as

$$\begin{aligned} H_{jk} &= \frac{(1 - \pi_j)\pi_k - (\pi_k - \pi_{jk})}{(1 - \pi_j)\pi_k} \\ &= \frac{\pi_{jk} - \pi_j\pi_k}{(1 - \pi_j)\pi_k}. \end{aligned}$$

The numerator of H_{jk} equals the covariance between binary variables, and the denominator is the maximum positive value of the covariance given the marginal distributions of the item scores. Let

$\text{cor}(X_j, X_k)$ denote the correlation between the responses on items j and k, and let $\text{cor}(X_j, X_k)_{\max}$ be the maximum correlation given the marginal distributions of items j and k. Then, H_{jk} can also be defined as

$$H_{jk} = \frac{\text{cor}(X_j, X_k)}{\text{cor}(X_j, X_k)_{\max}}.$$

Thus, the H_{jk} can be seen as a normed correlation coefficient—that is, a measure that takes into account that the correlation cannot reach the maximum value of 1, for items with different proportions, correct scores π_j and π_k.

From the pairwise H_{jk}, one can derive H_j for each item belonging to a scale, as well as the overall scale H. H_j indicates how well item j fits into a scale, and H indicates the strength of the scale. High values of H indicate a more correct ordering (i.e., fewer Guttman errors) of subjects on the latent trait.

The H_j for item j can be defined as follows:

$$H_j = 1 - \frac{\sum_{k \neq j} F_{jk}}{\sum_{k \neq j} E_{jk}} = \frac{\sum_{k \neq j} (E_{jk} - F_{jk})}{\sum_{k \neq j} E_{jk}},$$

as in Mokken (1971, p. 150). It can also be written as a weighted mean of H_{jk} coefficients, with weights equal to E_{jk}; that is,

$$H_j = \frac{\sum_{k \neq j} E_{jk} H_{jk}}{\sum_{k \neq j} E_{jk}}. \tag{1}$$

The scale H is defined as

$$H = 1 - \frac{\sum_j \sum_{k \neq j} F_{jk}}{\sum_j \sum_{k \neq j} E_{jk}} = \frac{\sum_j \sum_{k \neq j} (E_{jk} - F_{jk})}{\sum_j \sum_{k \neq j} E_{jk}}.$$

It can easily be verified that H can also be defined as a weighted mean of pairwise H_{jk} or item H_j,

$$H = \frac{\sum_j \sum_{k \neq j} E_{jk} H_{jk}}{\sum_j \sum_{k \neq j} E_{jk}} = \frac{\sum_j (\sum_{k \neq j} E_{jk}) H_j}{\sum_j \sum_{k \neq j} E_{jk}}, \tag{2}$$

where the weights are equal to E_{jk} and $\sum_{k \neq j} E_{jk}$, respectively.

Let $\text{cov}(X_j, X_k)$ denote the covariance of variables j and k. A set of J items is called a *Mokken scale* (Mokken, 1971, p. 184) if all items satisfy the following two conditions:

Condition 1 $\text{cov}(X_j, X_k) > 0$, for all $j \neq k$, and
Condition 2 $H_j \geq c$, for all j, where c is a user-defined constant between 0 and 1.

The first condition follows from the UD, LI, and M assumptions of the MH model (Mokken, 1971, p. 149; see also Holland & Rosebaum, 1986). This condition can also be restated as $\text{cor}(X_j, X_k) > 0$ or $H_{jk} > 0$. The second condition serves the practical purpose that only items with sufficient discrimination power are accepted into a scale (e.g., Sijtsma, 1998). More generally stated, the higher c, the more accurate the ordering of persons on the underlying trait θ. Mokken chose $c = 0.3$ in Condition 2, which, over the years, has been shown to be a good rule of thumb in measuring a single underlying dimension. Mokken also provided rules of thumb for interpreting the strength

of a scale (Mokken, 1971, p. 185). Hemker, Sijtsma, and Molenaar (1995) provided suitable values for c for finding the simulated dimensionality structure of item pools having different item characteristics.

For a set of items satisfying a Mokken scale, the following inequalities hold:

$$0 < \min(H_{jk}) \leq \min(H_j) \leq H \leq \max(H_j) \leq \max(H_{jk}) \leq 1 \tag{3}$$

(see Hemker et al., 1995; Mokken, 1971). This means, for example, that the lowest H_j in the scale is at least equal to the lowest H_{jk}. When searching for scales satisfying the Mokken scale conditions using a clustering procedure (sequential, hierarchical, or otherwise), one will generally see that H and also the H_j decrease when the number of items in the scale increases.

The H coefficient, as well as MSA in general, has been extended to polytomous items (Molenaar, 1982, 1991). The H coefficient for polytomous items also is the normed covariance between item pairs, and the interpretation of the coefficient remains the same (Hemker et al., 1995). Although the hierarchical procedures are discussed only for dichotomous items, they can be readily applied to item sets consisting of polytomous items.

Sequential Clustering

One way to find a Mokken scale is by checking whether an a priori selected set of items satisfies the two conditions described above. Items that do not fulfill these conditions should be removed. Another, more exploratory, approach involves applying a stepwise item selection procedure rather than specifying an a priori selected set. Such a stepwise procedure has been implemented in the program MSP (Molenaar & Sijtsma, 2000).

The sequential item selection procedure works as follows:

Step 1: Select the two items with the highest significantly positive H_{jk} in the sample.

Step 2: Compute H for all remaining items with respect to the already selected items, and select the item with the highest H that satisfies Conditions 1 and 2.

Step 3: Repeat Step 2 until no items remain that satisfy Conditions 1 and 2. If items remain, go to Step 1 to form another scale using the remaining items. If no more items remain, the entire procedure stops.

The default value for c is 0.3. The default α value used in the one-sided significance tests for H_{jk} and H_j is 0.05. To reduce the risk of capitalizing on chance, the level of significance is adjusted using a Bonferroni correction at each step. It is possible to change the default values for c and α, as well as to use other starting sets in Step 1. In the situation of a tie in Step 1 or 2, the item pair with the lowest π_j is selected.

It should be noted that MSP's item selection algorithm can also be regarded as a sequential clustering algorithm. The objects to be clustered are the items, and the matrix with the H_{jk} serves as proximity matrix between the items. The H serves as a proximity measure between a set of items forming a cluster or scale and the remaining single items that could be added to the cluster that is being built. Clusters are formed sequentially; that is, only after one cluster has been formed, item selection for a second cluster starts. The end result is a set of clusters, each consisting of two or more items and each forming a Mokken scale. Items that do not satisfy the scaling conditions for any scale in Step 2 are not entered in any cluster. In general, the higher the discrimination of the items in an item pool, the fewer nonscalable items there are.

Table 1

Lower Diagonal of H_{jk}-Matrix and Item Popularities π_j for Small Generated Example

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Item 1						
Item 2	0.87					
Item 3	0.32	0.31				
Item 4	0.01	–0.02	0.57			
Item 5	0.06	0.04	0.56	0.48		
Item 6	0.00	0.00	0.73	0.79	0.84	
π_j	0.68	0.38	0.50	0.75	0.62	0.27

A drawback of the sequential clustering procedure is, however, that it may yield a suboptimal solution in the sense that the true dimensionality of the item response data may not be found (van Abswoude et al., 2004). This phenomenon is illustrated by means of a small example consisting of the responses on six items, denoted Item 1, . . . , Item 6, and two independent latent traits, θ_1 and θ_2. Assume that Item 1 and Item 2 are strongly related to θ_1, Item 3 is weakly related to θ_1 and strongly related to θ_2, and Item 4, . . . , Item 6 are moderately related to θ_2 but not related to θ_1. The H_{jk} matrix and π_j values of these six items are presented in Table 1.

Using MSP's defaults settings, Step 1 will yield item pair Item 1-Item 2 as starting pair. Subsequently, Item 3 will be added to that cluster. The final first scale will consists of Item 1, Item 2, and Item 3, with $H = 0.46$. The second scale will contain the remaining three items-Item 4, Item 5, and Item 6, with $H = 0.65$. This two-cluster solution is suboptimal because there exists a solution that better reflects the true dimensionality of the item response data from the six items. This solution consists of two clusters containing Item 1-Item 2 and Item 3, Item 4, Item 5, Item 6, respectively. The H values for these scales are 0.87 and 0.63.

The correlation between the sum scores on subsets of these items was calculated; these subsets are item pair Item 1-Item 2 (Set 1), Item 3 (Set 2), and Item 4, Item 5, and Item 6 (Set 3). The Pearson product moment correlation coefficient (ρ) between the sum scores for Sets 1 and 2 equals 0.272; for Sets 1 and 3, $\rho = 0.022$; and for Sets 2 and 3, $\rho = 0.540$. Thus, these correlations indicate that there is no linear relationship between Sets 1 and 3 and that Item 3 seems to fit both in Sets 1 and 3, but the linear relationship with the items in Set 3 is strongest.

The problem of the sequential clustering procedure is, of course, that the starting set determines the solution. Once an item is selected in a cluster, it remains there, even if it fits better into a cluster that is formed later. The better solution would be obtained if the selection procedure starts with the second highest H_{jk} and then continues using MSP's default settings. The MSP software offers the possibility to specify other starting sets or extend the search procedure to allow for overlapping clusters, which means that it would have been possible to find the better solution. Unexperienced users may, however, not override the default settings and may therefore end up with suboptimal solutions.

Hierarchical Clustering

A possible solution to the problem associated with MSP's sequential item clustering method may be to switch to another type of clustering method. Hierarchical clustering analysis (HCA; e.g., Everitt et al., 2001) may be a useful method to find sets of items that form a Mokken scale.

The starting point of HCA is a data matrix containing proximities between separate objects. In this study, the separate objects are items, and their proximities are their pairwise H_{jk}. At each hierarchical step, the two objects that are most similar are joined. A joined pair is also called an object or cluster. This means that at any hierarchical step, two single items may be clustered to form one new cluster, a single item may be added to an existing cluster of items, or two clusters may be combined into a single larger cluster. This process continues until some previously stated criterion (i.e., a stopping rule) is met or until all items are in one single cluster.

Hierarchical clustering has been used before for dimensionality assessment in a classical test theory context (e.g., Bacon, 2001; Hunter, 1973; Reveille, 1979; Schweizer, 1991), as well as in a NIRT modeling context (Roussos et al., 1998). These studies have in common with each other and with this study that they use variants of product moment correlations coefficients as proximity metrics. For example, Reveille (1979) used Cronbach's alpha and the worst split-half (beta) coefficient to form clusters having high internal consistency reliability, and Roussos et al. (1998) used conditional correlations and conditional covariances to obtain clusters that are weakly locally independent. These studies differ from this study in the specification of minimal requirements on the clusters constructed with HCA, meaning that two items that correlate negatively could end up in the same cluster. New in this study's hierarchical clustering procedures is that clusters constructed with the HCA need to satisfy some minimal requirements, the Mokken scaling conditions.

Let O_v denote an object consisting of one or more items. The number of items in O_v is denoted by J_v, and the proximity between objects O_v and O_w is denoted by $H_{O_v O_w}$. The way the proximities are calculated will be explained later on. Hierarchical clustering of items uses the following steps:

Step 1: Join the two items j and k with the highest H_{jk} that satisfy the scaling conditions.

Step 2: Compute the $H_{O_v O_w}$ between all object pairs O_v and O_w, and join the object pair with the highest $H_{O_v O_w}$ as long as the stopping rule is not satisfied.

Step 3: Repeat Step 2 until no combination of two objects remain that satisfies the stopping rule.

When using HCA, individual items are not allocated to one cluster at the time, but multiple clusters may be formed simultaneously. This means that for the small example introduced in the previous section, HCA should be able to allocate Item 3 to the cluster it fits best in terms of dimensionality—that is, to the second instead of the first cluster.

Proximities

In this study, the performance of four types of agglomerative HCA methods was investigated: *complete linkage*, *average linkage*, *within-groups linkage*, and *scale linkage*. Although each of the methods uses the same three steps, they differ with respect to the definition of $H_{O_v O_w}$, or the proximity between clusters of items. The first three methods are available in the clustering routines of most statistical packages, such as SPSS (1998). The scale linkage method is a nonstandard method that is especially developed for the Mokken scaling problem.

A complete linkage method is obtained by defining the proximity between clusters as

$$H_{O_v O_w}^{complete} = \min(H_{jk}), \text{ where } j \in O_v \text{ and } k \in O_w.$$

In other words, at each step, those two objects—O_v and O_w—are joined for which the least similar pair of items has the highest proximity. Complete linkage is also known as the furthest neighbor method.

Average linkage, also known as the unweighted pair-group method of averages (Sokal & Michener, 1958) or between-groups linkage (Everitt et al., 2001), defines the proximity between objects as

$$H_{O_v O_w}^{average} = \frac{\sum_{j \epsilon O_v} \sum_{k \epsilon O_w} H_{jk}}{J_v J_w}.$$

As can be seen, $H_{O_v O_w}^{average}$ is the unweighted average of the bivariate H_{jk} *between* the items in object v and the items in object w. This measure of proximity therefore reflects the average distance of items belonging to different clusters.

Within-groups linkage defines the proximity of two objects, O_v and O_w, as the unweighted average of the H_{jk} of all items *within* O_v or O_w. In other words,

$$H_{O_v O_w}^{within} = \frac{\sum_j \sum_{k \neq j} H_{jk}}{(J_v + J_w)(J_v + J_w - 1)}, \text{ where } \{j, k\} \epsilon\ O_v \cup O_w.$$

The fourth method, scale linkage, is based on the scale H of the possible new object that is obtained by joining two objects; that is,

$$H_{O_v O_w}^{scale} = \frac{\sum_j \sum_{k \neq j} E_{jk} H_{jk}}{\sum_j \sum_{k \neq j} E_{jk}}, \text{ where } \{j, k\} \epsilon\ O_v \cup O_w.$$

Scale linkage may be regarded as the a hierarchical clustering variant of the sequential clustering procedure of the MSP package—that is, those objects are joined that together result in the largest possible scale H.

To be able to define stopping rules on the basis of the four types of proximity measures, it is important to get some idea on their meaning in the context of Mokken scaling. The first and last proximity measures have a direct interpretation in the context of Mokken scaling: $H_{O_v O_w}^{complete}$ is directly related to the minimal requirement for items belonging to the same scale (Condition 1; $H_{jk} > 0$), whereas $H_{O_v O_w}^{scale}$ is the overall summary measure of the quality of a scale (see equation (2)). The other two measures can be seen as proxies for Mokken scaling measures. As can be seen, $H_{O_v O_w}^{within}$ is an *unweighted* average of H_{jk} and can, therefore, be interpreted as an unweighted approximation of the overall scale H, which was a *weighted* coefficient of the H_{jk}s. Similarly, $H_{O_v O_w}^{average}$ can be seen as an unweighted proxy for H_j coefficients. More precisely, for each item belonging to object v, an unweighted item H_j could be computed indicating how well it fits into object w, $H_j^{O_w} = \sum_{k \epsilon O_w} H_{jk} / J_w$. The measure $H_{O_v O_w}^{average}$ equals the average of these $H_j^{O_w}$s: $H_{O_v O_w}^{average} = \sum_{j \epsilon O_v} H_j^{O_v} / J_v$.

Stopping Rule

The most natural point to stop a hierarchical clustering process is when the largest proximity drops below some minimum value; that is, stop the process of joining objects if $\max(H_{O_v O_w}) < c$. When applying the HCA in this study, it is important to choose the stopping rule in such way that it leads to clusters that satisfy the same scaling conditions as the ones required with sequential clustering.

As one of the reviewers of this article correctly pointed out, the choice of a stopping rule is a function of the research objective. For some objectives, it may be preferable to join subsets of items sensitive to highly correlated traits because joining also increases test reliability, and for other objectives, joining is undesirable. In MSA, constant c was incorporated (see scaling Condition 2) to allow such research objectives to be incorporated in scale construction. The stopping rules proposed here should be regarded as versions of this scaling condition that are adapted to the various HCA methods discussed in this article.

Defining a stopping rule based on $H^{complete}_{O_v O_w}$ is not straightforward because there is no direct relationship between the minimal H_{jk} of a scale and the overall homogeneity of the items (H_js). From Conditions 1 and 2 and equation (3), the requirement that $0 < H^{complete}_{O_v O_w} \leq c$ can be derived. It, therefore, seems reasonable to set the minimum value of $H^{complete}_{O_v O_w}$ somewhat larger than zero—say, $c^{complete} = 0.10$—to increase the quality of measurement.

Using the same line of reasoning, a hypothesis about the value of the stopping rules for the other methods can be derived using Conditions 1 and 2 and equation (3). The overall scale quality, $H^{scale}_{O_v O_w}$, and the unweighted overall scale quality, $H^{within}_{O_v O_w}$, should be at least as large as the value one would use for c in the sequential procedure (because $\min[H_j] \leq H$) and maybe somewhat larger, say, 0.40. Because of its relationship to the item H_j, the same minimum value seems to be reasonable for $H^{average}_{O_v O_w}$ (i.e., $c^{average} = c = 0.30$).

Rather than using *method-specific stopping rules*, it is also possible to use a more *general rule stopping rule*. This means that the methods still maximize different proximities at each clustering step, but for each method, the process of clustering stops when the same condition is no longer satisfied. As a general stopping rule, $H_j < c$, which is directly related to the Mokken scale conditions, can be used. In the simulation study, both method-specific and general stopping rulesare used.

Mokken Scale Conditions

In the sequential clustering algorithm, the conditions that define a Mokken scale are used as a stopping rule; that is, if the conditions are no longer fulfilled, the algorithm starts forming the next scale when any scalable items are left. The scales that are formed by means of HCA should satisfy the same Mokken scale conditions. The application of the conditions is, however, much more complicated within these clustering procedures.

The two Mokken scale conditions first need to be translated in such a way that they could be applied in HCA. For all objects O_v, one should check whether the following conditions hold:

$$H_{jk} > 0, \text{ for all } \{j, k\} \epsilon O_v,$$
$$H_j \geq c, \text{ for all items } j, \text{ where } j \epsilon O_v.$$

The above conditions could, as in the sequential clustering procedure, be checked within the clustering process, which amounts to using an additional stopping rule. This seems to be too strict for this study's HCA methods. Such as strategy could, for instance, impair the clustering of two objects just because a single item does not satisfy the Mokken scaling conditions while the remaining and possibly large set of items does satisfy the conditions. As a result, one would end up with many small clusters.

Alternatively, the Mokken scale conditions can be checked after clusters have been formed. A stopping rule is used to stop the clustering process at a certain number of clusters, and an additional step is added to the HCA method in which misfitting items are deleted from the scales. If there is more than one misfitting item in a scale, the items are deleted sequentially, where the worst item is deleted first. The worst misfitting item is that item j with the lowest H_j among the items with negative H_{jk}s, or the item with the lowest H_j when negative H_{jk}s do not exist. This is the procedure followed in the simulation study.

Simulation Study

Description of the Design

A simulation study was conducted to evaluate the performance of five clustering procedures: sequential, complete linkage, average linkage, within-groups linkage, and scale linkage.

Method-specific stopping rules were used for each clustering method. For sequential clustering, $c = 0.3$ and $c = 0$ were used. These conditions reflect two possible ways to conduct scale construction: quite restrictive ($c = 0.3$), as in MSP's default, and lowly restrictive ($c = 0.0$), as in Roussos et al. (1998) or Zhang and Stout (1999a). In hierarchical clustering, the four method-specific stopping rules had the values $c^{complete} = 0.1$, $c^{average} = 0.3$, and $c^{within} = c^{scale} = 0.4$. To facilitate the comparison between sequential clustering and hierarchical clustering, this study investigated the performance of $c = 0.3$ (for H_j) as a general stopping rule for the four hierarchical clustering methods.

Apart from the type of algorithm and the type of stopping rule, the data structure also was varied. More precisely, the size of the correlations between the traits (five conditions) and the number of items per dimension (three conditions) were varied. The number of latent traits was always three. Because we wanted to circumvent sampling fluctuation issues, extremely large sample sizes ($N = 100{,}000$), which approximate population data, were used. For a few cells, however, also the performance of the methods using a sample of $N = 200$ were investigated.

Three conditions with identical correlations for each pairs of traits ($\rho_{12} = \rho_{23} = \rho_{13}$) and two conditions with unequal correlations were used. The three conditions with identical correlations were 0.1, 0.4, and 0.7, representing weak, modest, and strong correlations, respectively. Because in most practical test applications, latent traits show dependencies (McDonald, 2000), a no-dependency condition ($\rho = 0.0$) was not included. The lower the ρ, the easier it should be to find the correct dimensionality. In the two situations with unequal correlations, the correlations were $\rho_{12} = \rho_{13} = 0.20$ and $\rho_{23} = 0.60$, as well as $\rho_{12} = \rho_{13} = 0.40$ and $\rho_{23} = 0.60$. Finding the correct dimensionality may be more difficult for unequal correlations than for equal correlations between latent traits. These different correlations are realistic in practical test applications.

The number of items per dimension was either 5 or 10; that is, each of the three traits may be represented by a small (5) or a large (10) number of items. Notation $[D : J_1; J_2; \ldots ; J_D]$ is used to reflect the structure of an item pool, where D equals the *simulated* number of traits and $J_d (d = 1, \ldots, D)$ equals the number of items sensitive to each trait. The three conditions used were $[3 : 5; 5; 5]$, $[3 : 10; 10; 10]$, and $[3 : 5; 5; 10]$, where the latter condition represents a situation with an unequal number of items per dimension. The situation with unequal item numbers was included because van Abswoude et al. (2004) encountered that the nonhierarchical clustering procedure DETECT (Zhang & Stout, 1999a) is less successful under such a condition.

The performance of the various procedures (methods and stopping rules) was evaluated by means of two criteria. The first criterion is whether the item selection procedure retrieves the true dimensionality of the problem or, in other words, whether items sensitive to the same latent trait are assigned to the same cluster. The second performance criterion is the overall fitness of the partition. This was quantified as the number of items that must be discarded because of misfit with respect to the two Mokken scaling conditions.

True Model of the Item Responses

The model used for generating item responses was a three-dimensional version of the five-parameter acceleration model (M5-PAM; van Abswoude et al., 2004; see also Sijtsma & van der Ark, 2001). This model was used because it has the flexibility to represent a large variety of nondecreasing item response functions (IRFs) that may be typical for nonparametric IRT models. In this study's M5-PAM model, the probability that subject i answers item j correctly, given his or her values on the three latent traits, equals

$$P(X_{ij} = 1|\theta_{i1}, \theta_{i2}, \theta_{i3}) = \gamma_j^{lo} + (\gamma_j^{up} - \gamma_j^{lo}) \left[\frac{\exp(\sum_{d=1}^{3} 1.7\alpha_{jd}\theta_{id} + \delta_j)}{1 + \exp(\sum_{d=1}^{3} 1.7\alpha_{jd}\theta_{id} + \delta_j)} \right]^{\xi_j}.$$

Table 2

Number of Clusters and Number of Items Per Cluster Obtained With Sequential Clustering for Simulated Data Sets

	Correlation Between Latent Traits				
Test Composition	0.1	0.4	0.7	0.2/0.6	0.4/0.6
			$c = 0.3$		
[3 : 5; 5; 5]	[3:5;5;5]	[2:11;4]	[1:15]	[2:11;4]	[2:11;4]
[3 : 5; 5; 10]	[3:5;5;10]	[2:16;4]	[1:20]	[2:16;4]	[1:19][a]
[3 : 10; 10; 10]	[3:10;10;10]	[1:30]	[1:30]	[2:21;9]	[1:30]
			$c = 0.0$		
[3 : 5; 5; 5]	[1:15]	[1:15]	[1:15]	[1:15]	[1:15]
[3 : 5; 5; 10]	[1:20]	[1:20]	[1:20]	[1:20]	[1:20]
[3 : 10; 10; 10]	[1:30]	[1:30]	[1:30]	[1:30]	[1:30]

a. One item excluded because of misfit with scale conditions.

Here, α_{jd} is the discrimination parameter of item j on trait d, and δ_j is frequently referred to as the difficulty parameter of item j. In M5-PAM, the slope and the location of the IRFs depend not only on the α_{jd} and δ_j parameters but also on γ_j^{lo}, γ_j^{up}, and ξ_j, which are the lower asymptote, the upper asymptote, and the acceleration parameter of the IRF of item j, respectively. The last parameter makes an IRF asymmetrical (see also Samejima, 2000).

Some items were assumed to be sensitive to a single latent trait (Zhang & Stout, 1999a, called this condition simple structure), whereas others were assumed to be strongly sensitive to one trait and less strong to the other two traits. Values for θ_{id} were drawn from a trivariate normal distribution with means of zero and correlations depending on the condition. The values of the other parameters were fixed: α_{jd} ranged between 1.50 and 2.25 for dominant traits and was set to 0.25 or 0.0 for nondominant traits, δ_j ranged from -1.5 to 1.5, γ_j^{lo} from 0.0 to 0.1, γ_j^{up} from 0.9 to 1.0, and ξ_j from 0.5 to 3.0. Item parameters were fixed to obtain items that are representative for true test situations. Another option would have been to generate parameters from certain distributions, but using this approach, it is difficult to obtain representative items.

Results

Sequential Clustering

Table 2 shows the detected dimensionality when using sequential clustering. The first column reports the true test composition. The remaining columns report the retrieved dimensionality for the five conditions related to the correlations between latent traits. Notation $[K : J_1; J_2; \ldots ; J_K]$ is used to represent the number of detected clusters (K) and the number of items per cluster (J_k).

The higher the correlations between traits ($\rho = 0.4$ or 0.7), the fewer the number of clusters detected. In these cases, items sensitive to differences on different latent traits were collected into one cluster. Clustering simulated data based on traits with varying correlations showed the same trend; that is, items that were sensitive to differences on highly correlating traits were collected in one large cluster. Varying the number of items led to approximately the same number of clusters. Changing the scaling conditions from $c = 0.3$ to $c = 0.0$ resulted into a single cluster containing all items.

Hemker et al. (1995) and van Abswoude et al. (2004) found similar results for the sequential clustering procedure. Increasing the correlations between latent traits, or decreasing c in Criterion 2, means that an increasing number of items satisfy the scaling conditions of the first cluster, and consequently, fewer items remain to be collected in the second or third cluster. If the sequential clustering was restricted using the scaling conditions (i.e., increase c to 0.3), depending on the type of data, the true dimensionality could be retrieved.

Given that K clusters are found (e.g., also when $K \neq D$), a solution may still be acceptable in the sense that items that are sensitive to differences on the same latent trait are in the same cluster. As can be seen in Table 2, sequential clustering did not always lead to acceptable solutions in this sense. The result [2 : 11; 4], where [3 : 5; 5; 5] and $\rho = 0.4$ were simulated, for example, means not only that items sensitive to differences on two different traits (θ_2 and θ_3) were selected into Cluster 1 but also that an item that mainly was sensitive to differences on θ_1 was selected into Cluster 1. The remaining items were selected into Cluster 2. Because objects were clustered one by one, the cluster that was formed first may be overrepresented compared to the cluster that was formed second.

In the main study, to get a global idea about the effect of ρ on the methods' success in dimensionality assessment, some values of the correlations between traits were investigated. For a few design conditions, it was investigated in more detail for which values of ρ the correct dimensionality was found, particularly for $\rho = 0.1,\ 0.2$, and 0.3 using a simulated data set having five item responses based on three latent traits and $c = 0.3$. For $\rho = 0.1$, the cluster solution was [3:5;5;5]; for $\rho = 0.2$, the cluster solution was [3:5;4;5]; and for $\rho = 0.3$, the cluster solution was [3:7/4/4]. The slashes indicate that the obtained clusters are sensitive to different traits. These additional results suggest that for $c = 0.3$, the true dimensionality is recovered up to a value of ρ that lies between 0.1 and 0.2.

Hierarchical Clustering

Three types of results are presented for the hierarchical clustering methods. First, to compare the performance of the five clustering methods, the number of clusters found and the number of items that were deleted sequentially due to misfit to the Mokken scaling conditions (denoted as the number of misfitting items) are reported. For the stopping rules, hypothesized c values were used. In general, the large sample was used, but for a few cells, the stability of the methods using small samples was also investigated. Second, to find out whether the correct rules of thumb were used for c, a different perspective was adopted: The ranges of the value of c that would have recovered the simulated number of latent traits are presented. These values may give more insight into the functioning of the four new proximities for different sets of items (for similar ranges for H_j using sequential clustering, see Hemker et al., 1995). Finally, it was investigated whether the graphical plots of the proximities at each hierarchical step contain information on the actual dimensionality.

Table 3 reports the number of clusters and the number of misfitting items (in parentheses). The first part shows the results using the general stopping rule $H_j < 0.3$ (all methods), and the second part shows the results of each hierarchical clustering method in combination with its method-specific stopping rule. The specific number of items in each cluster is not reported for clarity of Table 3. Before discussing the effects of the general and method-specific stopping rules, some general results are described.

As far as the number of clusters is concerned, the hierarchical clustering methods showed approximately the same pattern as sequential clustering: The number of clusters decreased when the correlations between latent traits increased and the restrictiveness of clustering decreased. Unlike sequential clustering, the number of clusters did, in some instances, change between conditions with different numbers of items per latent trait: The number of clusters decreased when the number of

Table 3

Number of Clusters (Number of Misfitting Items in parentheses) Obtained With Hierarchical Clustering for Simulated Data Sets

	Correlation Between Latent Traits				
Test Composition	0.1	0.4	0.7	0.2/0.6	0.4/0.6
General stopping rule					
	$H_j < 0.3$ (all methods)				
[3 : 5; 5; 5]	3	2	1	2	2
[3 : 5; 5; 10]	3	2	1	2	2
[3 : 10; 10; 10]	3	1	1	2	1
Method-specific stopping rules					
	$H_{O_v O_w}^{Complete} < 0.1$				
[3 : 5; 5; 5]	3	1(1)	1	2	1(1)
[3 : 5; 5; 10]	3	1(1)	1	2	1(1)
[3 : 10; 10; 10]	3	1	1	2	1
	$H_{O_v O_w}^{Average} < 0.3$				
[3 : 5; 5; 5]	5	6	6	5	5
[3 : 5; 5; 10]	5	5	4	4	6
[3 : 10; 10; 10]	4	4	4	4	4
	$H_{O_v O_w}^{Within} < 0.4$				
[3 : 5; 5; 5]	3	2	1	2	1(1)
[3 : 5; 5; 10]	2(5)	1(1)	1	1(4)	1(1)
[3 : 10; 10; 10]	3	1	1	1(2)	1
	$H_{O_v O_w}^{Scale} < 0.4$				
[3 : 5; 5; 5]	3	2	1	2	2
[3 : 5; 5; 10]	3	1(1)	1	2	1(1)
[3 : 10; 10; 10]	3	2	1	2	1

items decreased. This result, which is caused by minor differences in the parameter values when 5 or 10 items per trait were simulated, was found for some hierarchical methods (i.e., average linkage and within-groups linkage) using the proposed stopping rules.

Because hierarchical clustering forms clusters simultaneously instead of sequentially, HCA should collect items sensitive to the same trait into the same cluster. One can observe in Table 2 that especially the two-cluster solutions of sequential clustering were suboptimal. The results using HCA are better; that is, [2 : 10; 5] for $J = 15$, [2 : 15; 5] for $J = 20$, and [2 : 20; 10] for $J = 30$ (the number of items per cluster is not shown in Table 3). In HCA, one does not see that the items sensitive to differences on the same dominant trait were joined into different clusters.

The number of misfitting items as reported in Table 3 gives an indication of the quality of a solution. As can be seen, most misfitting items were found when ρ was low or moderate and for a small number of clusters. Misfitting items could, of course, be prevented by increasing the

restrictiveness of item clustering. However, the absence of misfitting items does not indicate that the true dimensionality was found because it might be a solution with too many clusters.

The *general stopping rule* ($H_j < 0.3$) was applied to all HCA methods. The results were the same for all of these methods (results are presented only once in Table 3). This means that the path item clustering initially took had no effect on the final solution. When comparing the results of the general stopping rule with sequential clustering, one may observe that the number of clusters was the same. As explained before, the specific items in each cluster were somewhat dissimilar. The improvement of hierarchical clustering over sequential clustering lies in the fact that, when using HCA, items sensitive to differences on the same latent trait were collected into the same cluster, whereas when using sequential clustering, this was not always the case.

The four HCA *methods*, in combination with their *method-specific stopping rules*, were not equally successful in finding the true dimensionality. In complete linkage, using the method-specific stopping rule for conditions with modest (i.e., $\rho = 0.4$) and high (i.e., $\rho = 0.7$) correlations between latent traits, the number of clusters found was less than the true number of latent traits. This suggests that the method-specific stopping rule (i.e., the minimum of H_{jk}) for complete linkage was too low to find the correct dimensionality. In within-groups linkage and, to a lesser extent, in scale linkage, which both use functions of the scale H, similar trends can be observed. Scale linkage performed better than within-groups linkage (see Table 3) because in scale linkage, a proximity measure was used that corrects for variations in item difficulties. In within-groups linkage, especially in the conditions with 10 items per latent trait, too few clusters were found. Using the average linkage method with the average linkage criterion, more clusters than the true number were found. For many item pools, this meant that a different cluster was generated for each possible combination of traits; that is, sets of items sensitive to differences on either the same latent trait or a same combination of latent traits were collected in one cluster. Seemingly, average linkage does not serve as good proxy for H_j. Correct recovery stopped at the following values of ρ: general stopping rule between 0.2 and 0.3, complete linkage between 0.2 and 0.3, within-groups linkage between 0.2 and 0.3, and scale linkage between 0.3 and 0.4 (not shown in Table 3). Thus, increasing ρ in smaller steps shows that the scale linkage method performed best. To sum up, scale linkage seems to be the only method that, in combination with the method-specific stopping rule, performs better than the general stopping rule.

These results may not hold up for small samples. Therefore, the success of the methods and the specific stopping rules was investigated for $N = 200$. Table 4 shows results of 10 replications for small (i.e., $N = 200$) and large (i.e., $N = 100{,}000$) sample sizes. The table presents two types of results. First, for each method, the number of times the correct three-cluster solution was obtained using the four hierarchical clustering methods is presented. Alternatively, the average number of clusters could have been presented over 10 replications. This, however, is not informative because it says nothing about whether and how often the correct number of clusters is obtained and whether this number of clusters corresponded with the true underlying dimensionality. Second, the standard error (SE) of the proximities when the correct clustering was found using the method-specific stopping rule is presented for small and large sample sizes. The SE is calculated using the proximity value obtained with $N = 100{,}000$ as the true score. Let Y_r denote the proximity value of an HCA method for replication r. In addition, let μ denote the average of the proximity for $N = 100{,}000$ (it is referred to in Greek notation because it approaches the population value). Then, the SE equals

$$SE = \sqrt{\frac{1}{10}\sum_{r=1}^{10}(Y_r - \mu)^2}. \tag{4}$$

Furthermore, the results obtained for $N = 100,000$ response patterns presented earlier (see Table 3) suggested that the correlation conditions $\rho = 0.7$, 0.2/0.6, and 0.4/0.6 might not yield the true

Table 4

Numbers of Correct Clusters (#) and the Proximities Standard Errors (SE) for Large and Small Samples

	Correlation Between Latent Traits							
Test Composition	0.1		0.2		0.3		0.4	
[3 : 5; 5; 5]	#	SE	#	SE	#	SE	#	SE
	Complete linkage							
$N = 200$	9	.053	9	.100	3	.099	0	—
$N = 100,000$	10	.003	10	.004	0	—	0	—
	Average linkage							
$N = 200$	1	.082	0	—	0	—	0	—
$N = 100,000$	1	.000	1	.000	0	—	0	—
	Within-groups linkage							
$N = 200$	3	.037	4	.094	0	—	0	—
$N = 100,000$	10	.000	10	.005	0	—	0	—
	Scale linkage							
$N = 200$	9	.039	9	.084	3	.081	1	—
$N = 100,000$	10	.000	10	.007	10	.005	0	—

Note. # represents number (out of 10) times the correct clusters were found using predefined stopping rules. SE represents the standard error of the statistics found for the correct solutions. — denotes the SE cannot be determined.

dimensionality. For that reason, the results for the conditions $\rho = 0.1, 0.2, 0.3$, and 0.4, which may find the true dimensionality, were present.

The results in Table 4 show that for small samples, the methods using the method-specific stopping rules worked less well than than those using large samples. As indicated by the standard errors of the proximities (see Table 4), the proximities in some small samples fell below and in others fell above the value of the predefined stopping rules. Thus, the stopping rules were not equally suitable for large and small sample sizes.

In the second type of results, a different perspective is adopted: Instead of fixing c, it was investigated for what ranges of c the correct dimensionality was obtained for the method-specific stopping rules. These ranges are presented in Table 5. All methods were able to find the correct number of clusters and joined the correct items. In general, the higher ρ, the more restrictive the scaling conditions need to be to find the true dimensionality.

The range of c strongly changes with the proximity used for clustering. Using H_{jk} as proximity (in complete linkage) is very attractive because a large range of c values leads to the correct result. Table 5 provides a confirmation that hypothesized values of c were too low for complete linkage, within-groups linkage, and scale linkage and too high for average linkage to find the correct solution.

As indicated by the preceding results, the use of the hypothesized stopping rules may not be the desirable approach to determine the number of dominant dimensions of a data set. The best moment to stop the clustering process depends on several characteristics of the data—that is, on the properties of the items (e.g., discrimination) and of the subjects (e.g., variability on the traits). Explorative

Table 5

Ranges of c Yielding the Correct Dimensionality Using Four Hierarchical Clustering Methods for Simulated Data Sets

	Correlation Between Latent Traits				
Test Composition	0.1	0.4	0.7	0.2/0.6	0.4/0.6
	Complete linkage				
[3 : 5; 5; 5]	.04−.41	.14−.41	.25−.42	.21−.41	.21−.42
[3 : 5; 5; 10]	.05−.41	.17−.41	.31−.42	.25−.42	.26−.41
[3 : 10; 10; 10]	.03−.41	.12−.41	.22−.43	.18−.42	.19−.42
	Average linkage				
[3 : 5; 5; 5]	.08−.28	.17−.28	.24−.28	.21−.28	.20−.28
[3 : 5; 5; 10]	.08−.28	.18−.28	.24−.28	.22−.28	.19−.27
[3 : 10; 10; 10]	.08−.27	.18−.28	.26−.28	.23−.27	.22−.28
	Within-groups linkage				
[3 : 5; 5; 5]	.36−.62	.45−.63	.54−.64	.51−.62	.43−.64
[3 : 5; 5; 10]	.41−.63	.51−.63	.59−.64	.55−.63	.56−.64
[3 : 10; 10; 10]	.39−.67	.49−.68	.59−.69	.55−.67	.55−.67
	Scale linkage				
[3 : 5; 5; 5]	.32−.60	.42−.61	.51−.62	.47−.60	.48−.61
[3 : 5; 5; 10]	.38−.61	.48−.62	.56−.62	.52−.61	.53−.61
[3 : 10; 10; 10]	.36−.64	.46−.65	.56−.66	.52−.64	.52−.65

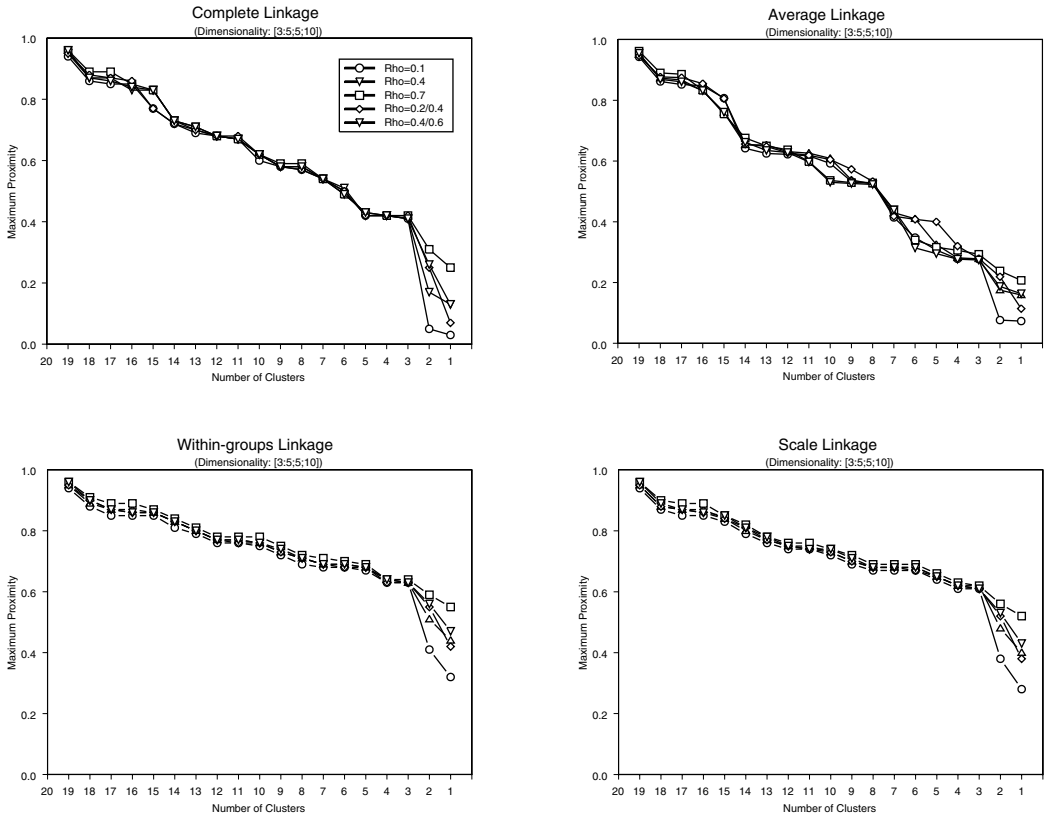
approaches that analyze the process of clustering may be more appropriate than confirmatory approaches that use predefined stopping rules in determining the moment to stop the clustering.

In the third type of results, scree plots of the proximities at each hierarchical step are presented. Figure 1 depicts the scree plots for item response data having unequal numbers of items per latent trait and five ρ conditions. Each plot depicts the drop in proximity (i.e., $\max(H_{O_v O_w})$ when two objects are joined into one cluster at each cluster step. One may note that Figures 1 and 2 depict not the drop in proximity when one item is added to a single scale but the drop when objects are joined. One or both of these objects may contain a single item. It is expected that the proximity function drops off when objects are joined because H generally decreases when more items are added to a scale. However, when objects sensitive to different traits are joined, one would expect a larger drop than when objects sensitive to the same latent trait are joined. Thus, with the sharp drop-off, the criterion aims to capture the drop in proximity that cannot be explained by the reduced correlation due to the numbers of items in the set(s) but by multidimensionality. One should note, however, that a sharp drop in proximity may not only occur for reasons of multidimensionality but also, for example, when there are one or two items having low discrimination in a set having high discrimination.

For equal numbers of items per latent trait, the complete linkage, within-groups linkage, and scale linkage plots indicate that the true dimensionality was found if the solution was used before a sharp drop in $\max(H_{O_v O_w})$ (at three clusters). The average linkage pattern was more difficult to interpret. Several larger drops can be seen, for example, for $\rho = 0.4$ and 0.7; a sharp drop in

Figure 1

Scree Plots of Number of Clusters and Maximum Proximity Using Complete Linkage, Average Linkage, Within-Groups Linkage, and Scale Linkage on Simulated Data Containing Unequal Numbers of Items Per Latent Trait and Five Levels of Correlations Between Latent Traits



$\max(H_{O_v O_w})$ can be observed at the six-cluster solution; and a less obvious drop can be observed for $\rho = 0.1$ at the three-cluster solution. Thus, for this method, it is more difficult to decide when to stop clustering.

The plots for equal numbers of items were very systematic in the sense that comparable objects were joined for each dimension. One may see in Figure 1 that $\max(H_{O_v O_w})$ does not change much for the first three steps, for example, because for each dimension, items with the same item characteristics (i.e., discrimination and difficulty) were joined. Because for unequal numbers of items per trait (i.e., true dimensionality: [3:5;5;10]) the items characteristics were not the same, the patterns were less systematic (see Figure 1). Although the pattern was less systematic, the general results for unequal numbers of items per latent traits remain the same as for equal numbers.

Empirical Example

The five clustering methods were compared using an empirical example. A dichotomized 15-item subset of the International Survey on Adult Crying (ISAC-A; Becht, Poortinga, & Vingerhoets, 2001), obtained for $N = 3{,}896$ subjects from 30 countries, was used. The questionnaire consists

of 54 items about common events and feelings that may induce crying. For clarity of presentation, this small subset of items was used. Cultural diversity in this analysis was ignored, and missing data were deleted listwise from the analysis.

In an empirical study, the theoretical constructs on which the test or questionnaire is based can be used to form hypotheses about the true dimensionality. From previous studies on the ISAC-A, two (Becht et al., 2001; Scheirs & Sijtsma, 2001) or three (Scheirs & Sijtsma, 2001) types of items can be distinguished: *distress*, representing emotions or situations that are unpleasant, for example, "I cry when having been humiliated or insulted" (Item 24); *sadness*, representing emotions or situations that are sad, for example, "I cry when I feel sad" (Item 1); and *joy*, representing emotions or situations that are happy, for example, "I cry when a movie or a television program has a happy ending" (Item 13). Becht et al. (2001) did not differentiate between the distress and sadness items. The H coefficient of these subtests was $H_{distress} = 0.50$ (8 items), $H_{sadness} = 0.47$ (3 items), and $H_{joy} = 0.34$ (4 items). The sum scores of the responses on the distress and sadness subsets were moderately related: Their Pearson product moment correlation was $\rho = 0.622$. The correlations between the responses on the distress and joy subtests (i.e., $\rho = 0.393$) and the sadness and joy subtests (i.e., $\rho = 0.357$) were much weaker.

The crying data were analyzed using the five clustering procedures. For sequential clustering, the same conditions were used as in the simulation study (i.e., $c = 0.3$, $\alpha = 0.05$). For hierarchical clustering, new stopping rules were specified using the ranges in Table 5; these are, $c^{complete} = 0.35$, $c^{average} = 0.25$, $c^{within} = 0.60$, and $c^{scale} = 0.55$. These new stopping rules are based on the results of the simulation study, which may not be representative for empirical data. Also, information about the process of clustering was obtained by making use of graphical analysis.

Results

Table 6 shows the results of the dimensionality analysis of the ISAC-A data. The first column shows the number of obtained clusters, the second column shows the theoretical constructs of each item per cluster, and the third column shows the $\max(H_{O_v O_w})$ of the newly clustered objects. The letters D, J, and S represent distress, joy, or sadness items, respectively. As can be seen from Table 6, no clustering method found the theoretical dimensionality. Sequential clustering resulted in a two-cluster solution, confirming that the responses on the distress and sadness items are substantially correlated. The hierarchical clustering methods yielded different results depending on the proximity and the stopping rule used for clustering. Using the new values for the stopping rules (denoted with an [a] in Table 6), the hierarchical methods found one large scale containing most distress items and one sadness item, as well as several smaller scales. One sadness item was always found in the cluster that mainly contained distress items because that item had the highest H_{jk}, with one of the distress items making this pair the starting set of all clustering methods. This shows a property of the items that is not detected when testing the homogeneity of the clusters confirmatory.

Scree plots of the ISAC-A data (see Figure 2) and complete linkage, average linkage, and within-groups linkage dendrograms (see Figure 3; a scale linkage dendrogram is not included because it cannot be replicated using SPSS and is therefore not readily available for applied researchers) may provide more reliable information about what solution to use. The dendrograms presented in Figure 3 depict the process by which items were joined, as well as the change in proximity between steps. The actual distances were rescaled to numbers between 0 and 0.25, preserving the ratio of differences between steps (SPSS, 1998).

In Figure 2, complete linkage had the most informative plot for determining the number of underlying traits: A clear cutoff point can be found at the three-cluster solution (see also Figure 3). Going from four to three clusters did not result in a large decrease in $\max(H_{O_v O_w})$; therefore, the

Table 6

Number of Clusters, Item Number and Item Type, and Maximum Proximity Using Five Clustering Methods for International Survey on Adult Crying (ISAC-A) Data

# Clusters	Theory	$\max(H_{O_v O_w})$
Sequential clustering ($c = 0.3$)		
2	[D6,D7,D8,D10,D11,D12,D13,D14,J15,S1,S5,S9] [J2,J3,J4]	—
Complete linkage		
4[a]	[J3,J4] [D14,D15,S5] [D7,D8,D10,D11,D13,D6D12,S9] [S1]	0.37
3	[J2,J3,J4] [D14,J15,S1,5S] [D7,D8,D10,D11,D13,D6,D12,S9]	0.32
2[b]	[J2,J3,J4][D6,D7,D8,D10,D11,D12,D13,D14,J15,S1,S5,S9]	0.19
Average linkage		
4[a]	[J3,J4] [D7,D8,D10,D6,D12,J15,J2,S1,S9] [D14,S5] [D11,D13]	0.28
3	[J3,J4][D14,S5][D11,D13,D7,D8,D10,D6,D12,J15,J2,S1,S9]	0.24
2	[J3,J4,D14,S5][D11,D13,D7,D8,D10,D6,D12,J15,J2,S1,S9]	0.23
Within-groups linkage		
10[a]	[J3][J4][D6,D7][D8,D10,D13,D12,S9][D11][J15][D12][J2][S1][S5]	0.61
...	...	...
4	[J3,J4][J15][D6,D7,D8,D10,D11,D12,D13,D14,S1,S5,S9][J2]	0.49
3[b]	[J3,J4][J15,D6,D7,D8,D10,D11,D12,D13,D14,S1,S5,S9][J2]	0.48
2	[J3,J4][J2,J15,D6,D7,D8,D10,D11,D12,D13,D14,S1,S5,S9]	0.45
Scale linkage		
8[a]	[J2] [J3] [D14] [J4] [J15] [S1] [S5] [D7,D8,D10,D11,D13,D6,D12,S9]	0.56
...	...	...
4	[J3,J4][D6,D7,D8,D10,D11,D12,D13,J15,S1,S5,S9][D14][J2]	0.49
3[b]	[J3,J4][D6,D7,D8,D10,D11,D12,D13,D14,J15,S1,S5,S9][J2]	0.47
2	[J3,J4][D6,D7,D8,D10,D11,D12,D13,D14,J15,J2,S1,S5,S9]	0.44

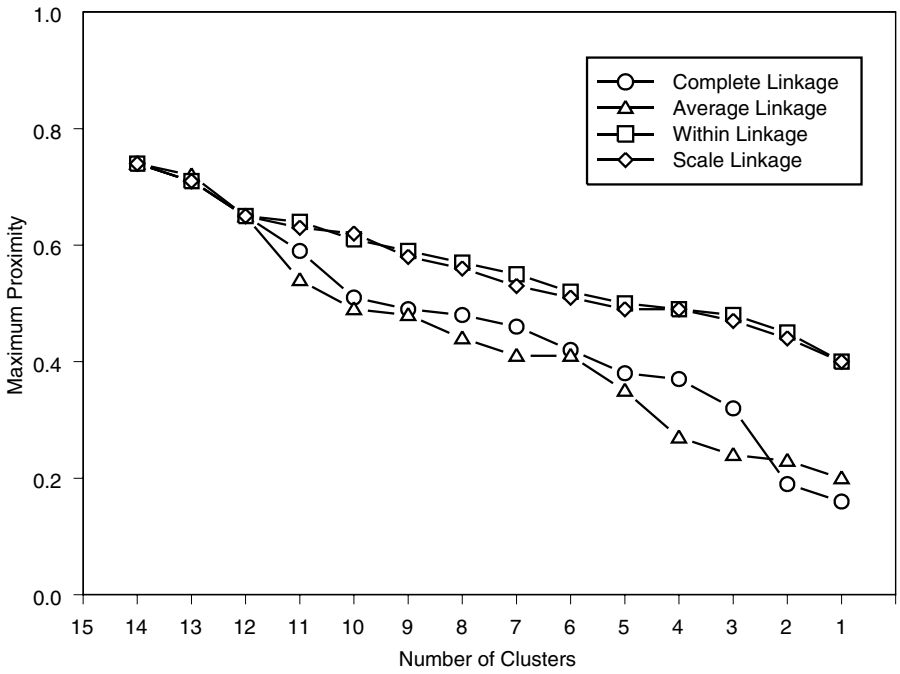
Note. The solutions still include misfitting items.
a. Result for hypothesized values of the stopping rules.
b. Result for general stopping rule, $H_j < 0.3$.

three-cluster solution was the dimensionality according to complete linkage. Within-groups linkage and scale linkage showed only minor drops in $\max(H_{O_v O_w})$ at $K = 3$ and were therefore less clear with respect to the dimensionality of these items. The pattern of average linkage was difficult to interpret: There were several small decreases but no clear cutoff point.

The relationship between the scores operationalized by the methods and correlations between the sets can help to interpret the results of the scree plots. Complete linkage joins the two objects having the highest minimum H_{jk} at each clustering step; thus, clustering is based on the relationship between each pair of variables. The other methods use proxies for H_j or H that, in general, represent the relationship between more than two variables. Thus, the other methods may average out the differences that complete linkage focuses on. Complete linkage could therefore depict a sharp drop-off where the other methods could not.

The lack of a clear cutoff point for all methods except complete linkage can be explained by the relatively high correlations between the sets. In particular, the sets measured common variables;

Figure 2
Scree Plots of Number of Clusters and Maximum Proximity Using Complete Linkage, Average Linkage, Within-Groups Linkage, and Scale Linkage of International Survey on Adult Crying (ISAC-A) Data



therefore, the proximity did not drop much when items sensitive to these variables were clustered. Although a sharp drop-off was observed for complete linkage at $K = 3$, the method's plot also indicates that the sets were moderately correlated. The proximities' values at $K = 2$ and $K = 1$, which were still quite high (probably even significantly positive for this sample size; see Condition 1), indicated this.

Figure 3 illustrates that the process of clustering is different in the various methods. For example, sequential clustering, complete linkage, and scale linkage (not shown) identify joy as a separate cluster in an early stage of the clustering process, whereas average linkage and within-groups linkage do not. To summarize the results of Figure 3, complete linkage was most useful as a method for determining the number of underlying variables, and complete linkage and scale linkage methods were successful for correctly identifying unidimensional sets.

When looking at the $K = 2$ solutions in Table 6, it can be seen that complete linkage leads to the same result as sequential clustering. One may note that the $K = 2$ solution of complete linkage combines two clusters of the $K = 3$ solution (i.e., the best solution according to the scree plot). Even though the results for $K = 2$ were the same, complete linkage should be preferred to sequential clustering because with complete linkage, more certain statements can be made about the dimensionality of the items. In sequential clustering, it is unknown whether the two clusters were combined (i.e., into the largest cluster) because this yields the strongest Mokken scale or because of the sequential nature of the item selection procedure (i.e., forming clusters one at the time). In hierarchical clustering, such information is present because $H_{O_v O_w}$, for all combinations of objects, is calculated, and only the two objects that maximized $H_{O_v O_w}$ are combined.

Figure 3

Dendrograms of Complete Linkage, Average Linkage, and Within-Groups Linkage of International Survey on Adult Crying (ISAC-A) Data

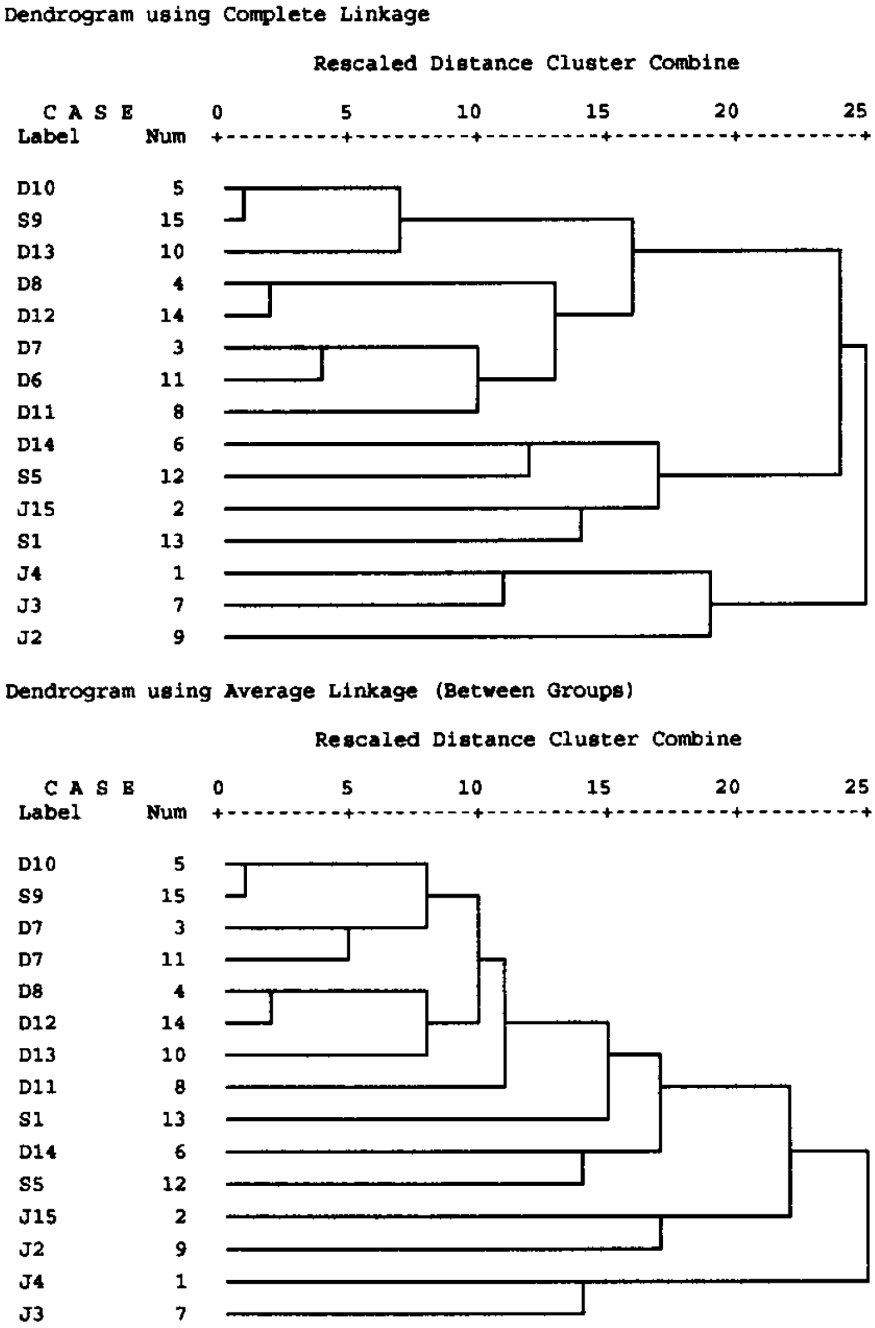
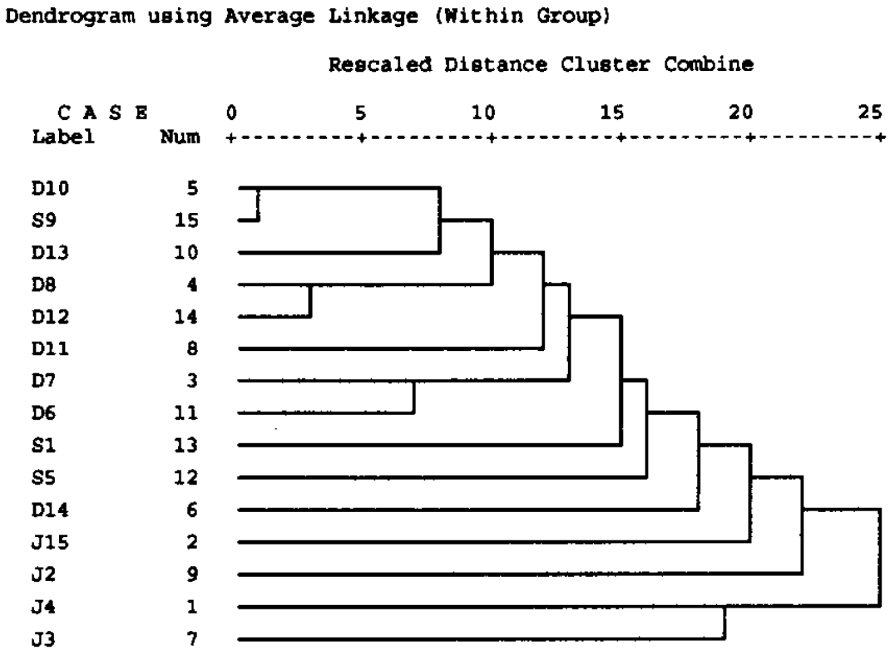


Figure 3
(continued)



Using complete linkage and interpreting its graphically depicted result, the three-cluster solution should be considered to be the best (e.g., clear results in scree plot, satisfies Mokken scale conditions). Based on substantive grounds or on the correlations between the sets, however, one might prefer the two-cluster solution with the lower within-cluster homogeneity.

Discussion

In this study, it was investigated whether hierarchical clustering improves sequential clustering that is the standard in Mokken scale analysis. The simulation study showed that all four HCA methods (i.e., complete linkage, average linkage, within-groups linkage, and scale linkage) were able to find the true dimensionality. However, the success of the hierarchical methods depended on the stopping rules of the clustering process. Scale linkage was the promising hierarchical clustering method because it found the dominant dimensionality for most data. Complete linkage also was promising because for this method, a large range of $c^{complete}$ values lead to the true dimensionality, and because scree plots were interpretable for this method. Finally, the general stopping rule used in combination with each of the HCA methods, seems to work well. This was to be expected because that rule is most closely related to the Mokken scaling conditions. In the empirical study, not all methods yielded the same results. Here, the underlying traits seemed to be substantially correlated, and again, only complete linkage displayed an interpretable scree plot.

There are two reasons why scale linkage and complete linkage are an improvement over sequential clustering. First, these methods yield better results, meaning that items sensitive to the same

latent trait are more often collected in one cluster using these methods than those methods using sequential clustering. Second, the clustering process is more informative: It shows which objects are joined at what step, and it shows the relative difference in maximum proximity between clustering steps. Using graphical methods (i.e., scree plots or dendrograms), the "best" dimensionality for a particular set of items can be found; that is, the solution before a sharp drop in maximum proximity is seen. One should be aware, however, that the best solution in terms of relative difference might not satisfy the Mokken scaling conditions. In addition, the presence of a sharp drop may depend on characteristics of the data. For example, for highly correlated latent traits, no sharp drop can be expected. Thus, the decision about which dimensionality to use should also be based on H_j and H values of clusters. The process information of sequential clustering is less informative. It does not inform whether any subclusters (i.e., these are the objects contained in a cluster) exist. Using hierarchical clustering methods, there is information about the subclusters and their scalability, and this information can be used to find the true dimensionality.

Sample fluctuation issues were ignored for the largest part of this article by using very large sample sizes. For small sample sizes (say, $N = 200$), it may occur that the H_{jk} for some low-discrimination items is negative due to sample fluctuation rather than because they are sensitive to different latent traits. The same line of reasoning goes for values of H_j near c. For a few cells of the simulation study, this study investigated the stability of the results for $N = 200$ and found confirmations of expectations. In future research, it may be worthwhile to address the impact of sample fluctuation on dimensionality assessment more explicitly.

As is common in sequential item selection, when creating scales, it is wise to make use of other available information. More specifically, the clustering process as presented in a dendrogram, substantive information, and methods used to search for specific violations of the MH model can provide additional information about the dominant underlying dimensionality of data.

Rather than using a hierarchical procedure as an alternative to sequential clustering, a nonhierarchical procedure in which an overall criterion is optimized could also have been used. For example, the H_j could be calculated for all items j and a given number of k clusters, and item j can be assigned to that cluster that maximizes H_j. The advantage of such a procedure is that not only objects but also items within objects are compared. In future research, nonhierarchical procedures will be addressed.

References

Bacon, D. R. (2001). An evaluation of cluster analytic approaches to initial model specification. *Structural Equation Modelling*, *8*(3), 379-429.

Becht, M. C., Poortinga, Y. H., & Vingerhoets, A. J. J. M. (2001). Crying across countries. In A. J. J. M. Vingerhoets & R. R. Cornelius (Eds.), *Adult crying: A biopsychosocial approach* (pp. 135-158). Hove, UK: Brunner-Routledge.

Douglas, J., Kim, H. R., Roussos, L., Stout, W. F., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992 administrations*. Newton, PA: LSAT.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*. London: Arnold/Oxford 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. *The Annals of Statistics*, *14*, 1523-1543.

Hunter, J. E. (1973). Methods of ordering the correlation matrix to facilitate visual inspection and preliminary cluster analysis. *Journal of Educational Measurement*, *10*, 51-61.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*, 109-132.

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. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99-114.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: de Gruyter.

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

Molenaar, I. W. (1991). A weighted loevinger H-coefficient extending the 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 [Software manual]. Groningen, the Netherlands: iec ProGAMMA.

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

Reveille, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*, 57-74.

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.

Scheirs, J. G. M., & Sijtsma, K. (2001). The study of crying: Some methodological considerations and a comparison of methods for analyzing questionnaires. In J. J. M. Vingerhoets & R. R. Cornelius (Eds.), *Adult crying: A biopsychosocial approach* (pp. 277-298). Hove, UK: Brunner-Routledge.

Schweizer, K. (1991). Classifying variables on the basis of disaggregate correlations. *Multivariate Behavioral Research*, *26*, 435-455.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*, 3-31.

Sijtsma, K., & van der Ark, L. A. (2001). Progress in NIRT analysis of polytomous item scores: Dilemmas and practical solutions. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 297-318). New York: Springer.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409-1438.

SPSS. (1998). *SPSSX user's guide*. New York: McGraw-Hill.

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study on test data dimensionality procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3-24.

Zhang, J., & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*, 129-152.

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

Acknowledgments

The authors would like to thank Klaas Sijtsma and two anonymous reviewers for their useful comments on the article. They also thank Marleen Becht and Ad Vingerhoets for their permission to use their data on adult crying.

Author's Address

Address correspondence to Jeroen Vermunt, Department of Methodology and Statistics, P.O. Box 90153, 5000 LE Tilburg, the Netherlands; e-mail: J.K.Vermunt@uvt.nl.