A comparative study of test dimensionality assessment procedures under nonparametric IRT models

van Abswoude, A.A.H.; van der Ark, L.A.; Sijtsma, K.

Published in:
Applied Psychological Measurement

Publication date:
2004

A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models

Alexandra A. H. van Abswoude, L. Andries van der Ark and Klaas Sijtsma

Applied Psychological Measurement 2004; 28; 3
DOI: 10.1177/0146621603259277

A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models

Alexandra A. H. van Abswoude, L. Andries van der Ark, and Klaas Sijtsma
Tilburg University, The Netherlands

In this article, an overview of nonparametric item response theory methods for determining the dimensionality of item response data is provided. Four methods were considered: MSP, DETECT, HCA/CCPROX, and DIMTEST. First, the methods were compared theoretically. Second, a simulation study was done to compare the effectiveness of MSP, DETECT, and HCA/CCPROX using the default settings of each program in finding a simulated dimensional structure of a matrix of item response data. In several design cells, the methods that use covariances conditional on the latent trait (DETECT and HCA/CCPROX) were superior in finding the simulated structure to the method that used normed unconditional covariances (MSP). Third, the correctness of the decision of accepting or rejecting unidimensionality based on the statistics used in DETECT and DIMTEST was considered. This decision did not always reflect the true dimensionality of the item pool. *Index terms: DETECT software and method, dimensionality of item response data, DIMTEST software and method, HCA/CCPROX software and method, MSP software and method, multidimensional item response data, nonparametric item response theory, unidimensional item response data.*

Introduction

Although it can be argued that test performance often is simultaneously governed by several latent traits, most researchers seem to agree that a test or a questionnaire should preferably measure only one dominant latent trait. This is reflected by the existence of many unidimensional item response theory (IRT) models and only a few multidimensional IRT models (e.g., Kelderman & Rijkes, 1994; Reckase, 1997). There are at least two reasons why unidimensional measurement is preferred.

First, when test data measure one latent trait, a single score can be assigned to each examinee, and the interpretation of test performance is unambiguous. Also, when a measurement practitioner intends to measure multiple latent traits, it can be argued that he or she should construct a unidimensional test for each trait separately. When items measuring different traits are part of the same test (e.g., when some items are sensitive to vocabulary and others are sensitive to verbal comprehension), this line of reasoning would stipulate that the test be split into two unidimensional subtests and that examinees obtain separate scores on each. Note that if one summary score were assigned based on both item types, it would be unclear to what degree a latent trait influenced the test score

Applied Psychological Measurement, Vol. 28 No. 1, January 2004, 3–24
DOI: 10.1177/0146621603259277

of a particular examinee because one ability could have compensated for the other, also depending on the strength of their mutual relationship.

Second, due to the larger number of parameters, the estimation of multidimensional IRT models is more complicated than the estimation of unidimensional IRT models (e.g., see Béguin & Glas, 2001, who used Markov chain Monte Carlo techniques for estimating a multidimensional normal ogive model). Using the simpler unidimensional IRT models instead may be an attractive option, particularly after an item clustering method has been applied to the data to determine their dimensionality. Then, a unidimensional IRT model can be fitted to the items loading on a particular latent trait, and this may be repeated for each latent trait.

Traditionally, the dimensionality of responses from a set of dichotomous items was determined using linear factor analysis. It is well known that "difficulty factors" may arise (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar & Stout, 1993; see Miecskowski et al., 1993, for an example) when items vary widely in difficulty, and correlations are based on binary item scores. Other problems may arise when tetrachoric correlations are used to correct for the extreme discreteness of the binary item scores. One problem is that the tetrachoric correlation matrix may not be positive definite (Knol & Berger, 1991; Lord & Novick, 1968, p. 349). Another problem is that tetrachoric correlations estimate a correlation based on hypothesized normal variables when, in fact, only binary scores were observed, and normality thus may be an invalid assumption. An alternative may be nonlinear factor analysis, but Hattie et al. (1996) found that nonlinear factor models were not as effective in discriminating between unidimensional and multidimensional data sets as their linear counterparts.

An alternative to factor analysis is nonparametric item response theory (NIRT), which is central in this study. NIRT uses a nonlinear model for the relation between binary correct/incorrect item scores and a continuous latent trait and has the advantage that it can be applied directly to the binary item scores. This means that tetrachoric correlations are not necessary. The purpose of this study was to investigate the effectiveness of three methods used for retrieving the dimensionality of binary item score data, which are based on NIRT and use covariances between binary item scores. Three methods are considered here as they exist "off the shelf": MSP (Hemker, Sijtsma, & Molenaar, 1995; Molenaar & Sijtsma, 2000), DETECT (Kim, 1994; Zhang & Stout, 1999a, 1999b), and HCA/CCPROX (Roussos, 1992; Roussos, Stout, & Marden, 1998). In addition, the statistical procedure DIMTEST (Nandakumar & Stout, 1993; Stout, 1987; Stout, Douglas, Junker, & Roussos, 1993; Stout, Goodwin Froelich, & Gao, 2001) was used for testing hypotheses about the dimensionality of item response data, and results were compared to the results of the other methods.

Nonparametric IRT

Strictly and Essentially Unidimensional Models

Strictly unidimensional models. Let $\mathbf{X} = (X_1, \ldots, X_J)$ be the vector of J binary scored item variables, and let $\mathbf{x} = (x_1, \ldots, x_J)$ be the realization of $\mathbf{X}$. Score 1 indicates a correct answer and score 0 an incorrect answer. The probability of an item score of one depends on one latent trait θ, and is denoted $P_j(\theta)$. This is the unidimensionality (UD) assumption. Probability $P_j(\theta)$ is the item response function (IRF). Furthermore, local independence (LI) is assumed, which is defined as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta). \quad (1)$$

Assumption LI means that, given any value of θ, the responses of an individual to the J items are statistically independent. Assumptions UD and LI together do not imply falsifiable consequences

on the observed data (Holland & Rosenbaum, 1986; Junker, 1993). For this purpose, restrictions are needed on the IRFs. For example, let θ_a and θ_b be the latent trait values of examinees a and b; then the monotonicity assumption (M) states that

$$P_j(\theta_a) \leq P_j(\theta_b), \text{ whenever } \theta_a < \theta_b, \text{ for } j = 1, \ldots, J.$$

Assumption M means that the IRFs are monotone nondecreasing in θ. The assumptions of UD, LI, and M together define the model of monotone homogeneity (Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002, chaps. 2-5). The model of monotone homogeneity is an NIRT model that implies the stochastic ordering of θ by the total test score, $X_+ = \Sigma X_j$ (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997).

Essentially unidimensional models. Stout (1990; see also Junker, 1993) defined the dimensionality of item response data in terms of the minimum number of traits necessary to achieve LI and M. In essentially unidimensional models, however, the assumptions of LI and M are relaxed to essential independence and weak monotonicity, respectively. Stout assumed that test performance is governed by a dominant latent trait and several nuisance latent traits. Following this idea, a vector $\boldsymbol{\theta} = (\theta, \theta_1, \ldots, \theta_W)$ represents the dominant θ and W nuisance traits. Based on large sample theory, *essential independence* (EI) (Stout, 1990) states that

$$\frac{2}{J(J-1)} \sum_{1 \leq j < k \leq J} |\text{Cov}(X_j, X_k | \boldsymbol{\theta} = \theta)| \to 0, \text{ for } J \to \infty;$$

also see McDonald (1982) and Holland and Rosenbaum (1986). For finite J, the analog to the large sample version of EI is that $\text{Cov}(X_j, X_k|\boldsymbol{\theta}) \approx 0$, which is mathematically idealized to *weak local independence* (weak LI) or, equivalently, *pairwise local independence;* that is,

$$\text{Cov}(X_j, X_k | \boldsymbol{\theta} = \theta) = 0, \text{ for all } \theta, \text{ and for all } 1 \leq j < k \leq J \tag{2}$$

(Stout et al., 1996; Zhang & Stout, 1999a). Note that weak LI (equation (2)) is an implication of LI (equation (1)) but not the other way around. In practice, weak LI may be used to investigate LI (Stout, 1990).

Weak monotonicity means that the average of J IRFs is an increasing function of $\boldsymbol{\theta}$ but leaves the individual IRFs unrestricted within the confines of this condition on the mean; that is,

$$J^{-1} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}_a) \leq J^{-1} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}_b), \text{ whenever } \boldsymbol{\theta}_a < \boldsymbol{\theta}_b, \text{ coordinatewise.}$$

Thus, the strictly unidimensional model has a stronger independence assumption and a stronger monotonicity assumption than the essentially unidimensional model.

Discussion of the models. Although both have different points of departure, the essentially and strictly unidimensional IRT models both imply weak LI. For analyzing empirical data, both types of models may use this property. For example, in the strictly unidimensional Rasch model, the LI assumption is investigated for empirical test data using statistical tests based on weak local independence (Molenaar, 1983; see also Glas & Verhelst, 1995). The most pronounced difference between the strictly and essentially unidimensional NIRT model discussed here is the investigation of the dimensionality of the responses to a set of items. Item selection based on strictly unidimensional models aims at finding one or more homogeneous (i.e., measuring one θ each) clusters, using observable consequences of the model of monotone homogeneity, in particular, of assumption M. Item selection based on essentially unidimensional models aims at finding clusters of items sensitive to one dominant trait each, using observable consequences of weak LI. These differences are explained in the next sections in more detail.

Methods for Investigating Dimensionality

MSP

Let a set of items consist of J dichotomous items, and let a unidimensional cluster of items consist of L items ($j = 1, \ldots, L; L \leq J$). The Mokken Scale Analysis for Polytomous Items computer program (MSP5 for Windows, or MSP for short; Molenaar & Sijtsma, 2000) uses scalability coefficient H (Loevinger, 1948; Mokken, 1971) as the criterion for selecting items that yield a unidimensional cluster. For items j and k, the H coefficient is defined as the ratio of the covariance between items j and k and their maximum covariance given the marginal distributions of the items; that is,

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

Thus, H_{jk} is the normed covariance of an item pair. The scalability coefficient of a single item j with respect to the other $L - 1$ items selected into a cluster is defined as

$$H_j = \frac{\sum_{k \neq j} \text{Cov}\,(X_j, X_k)}{\sum_{k \neq j} \text{Cov}\,(X_j, X_k)_{\max}}.$$

The item scalability coefficient H_j can be interpreted as an index for the slope of the IRF of item j. For example, under the two-parameter logistic model (2-PLM) (e.g., Birnbaum, 1968), fixing the distribution of θ and also the 2-PLM location parameters of the IRFs, the H_js are an increasing function of the slope parameters (Mokken, Lewis, & Sijtsma, 1986).

Finally, for a set of L items, the scalability coefficient H is a weighted average of the item H_js, with positive weights depending on the marginals. Let π_j be the proportion correct on item j, and write $\text{Cov}(X_j, X_k)_{max} = \pi^{(0)}_{jk}$. Note that $\pi^{(0)}_{jk} = \pi_j(1 - \pi_k)$ if $\pi_j \leq \pi_k$, and $\pi^{(0)}_{jk} = \pi_k(1 - \pi_j)$ if $\pi_k < \pi_j$. Mokken (1971, p. 152) writes coefficient H as

$$H = \frac{\sum_{j=1}^{L-1} \sum_{k=j+1}^{L} \pi^{(0)}_{jk} H_j}{\sum_{j=1}^{L-1} \sum_{k=j+1}^{L} \pi^{(0)}_{jk}}. \tag{3}$$

Because fixed π_js also imply fixed $\pi^{(0)}_{jk}$s, an increase of the H_js causes an increase of H. Under UD, LI, and M, it can be shown that $0 \leq H \leq 1$ (Mokken, 1971, p. 150). Given UD, LI, and M, the value of $H = 0$ means that the IRFs of at least $(L - 1)$ items are constant functions of θ, and $H = 1$ means that there are no Guttman errors (given that $\pi_j \leq \pi_k$, a Guttman error is defined as $X_j = 1$ and $X_k = 0$); see Mokken (1971, p. 150) for further elaboration. Mokken (p. 184) defined a scale as follows:

DEFINITION: A cluster of items is a *Mokken* scale if

$$Cov(X_j, X_k) > 0, \text{ for all item pairs } (j, k; j \neq k); \text{ and} \tag{4}$$

$$H_j \geq c > 0, \text{ for all items } j, \tag{5}$$

where c is a positive lower bound of H_j, which is user specified. The higher c, the more restrictive item selection is with respect to the discrimination of the items. A high c means good item discrimination and accurate person ordering using X_+ (see also Sijtsma & Molenaar, 2002, p. 68).

MSP uses a sequential bottom-up item clustering procedure to partition a multidimensional set of items into clusters of items that each constitute a Mokken scale. The default start set is the item pair in the pool with the highest significant positive H_{jk} (for other possibilities, see Molenaar & Sijtsma, 2000, chap. 5). The second step is the selection of an item from the remaining items that satisfies equations (4) and (5) with respect to the previously selected items and maximizes the common H of the already selected items and the newly selected item. In the next steps, items are added to the already selected cluster using the same procedure. A scale has been completed when no more items remain that satisfy equations (4) and (5). If items remain unselected, subsequent clusters of items may be selected as described for the first cluster. The procedure stops when no more items remain that satisfy equations (4) and (5). For more details about the item selection procedure, see Hemker et al. (1995) and Molenaar and Sijtsma (2000).

Additional remarks. First, by selecting Mokken scales using scaling condition $H_j \geq c$, the dimensionality of the data is implicitly investigated as well (see Hemker et al., 1995). Consider the following idealized situation. Assume that some items are driven by θ_1 and other items by θ_2 and that these traits are correlated. Notice that, for the entire set of items, an IRF is the regression of X_j on a composite of these two θs and that H_j expresses the strength of this relationship. Finally, assume that the relationship of the items driven by θ_1 with θ_1 is stronger than that of the items driven by θ_2 with θ_2. The rest score, $R_{(-j)} = X_+ - X_j$, estimates the latent trait composite, and the regression of item j on $R_{(-j)}$ is given by $P[X_j = 1|R_{(-j)}]$. Based on these assumptions, in general, the regression of items driven by θ_1 on $R_{(-j)}$ is steeper (higher H_j) than that of the items driven by θ_2 (lower H_j).

Suppose that the item pair selected first is driven by θ_1, then a conveniently chosen c value selects the other items sensitive to θ_1 into the first cluster because their H_js with respect to the already selected items are greater than those of items sensitive to θ_2. If these latter items have $H_j s < c$, they remain unselected and the first item cluster is completed. Because the remaining items are driven by θ_2, rest score $R_{(-j)}$ based on these items estimates θ_2, and the regression, $P[X_j = 1|R_{(-j)}]$, is steeper, resulting in higher H_js. If these H_js exceed lower bound c, then a second cluster consisting of items sensitive to θ_2 is selected.

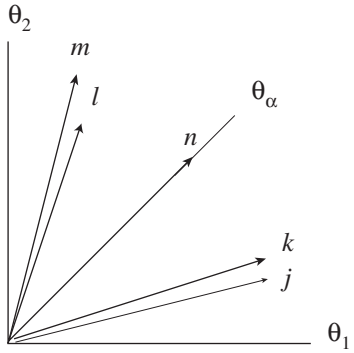
The choice of lower bound c affects the cluster composition. A low c value may result in clusters that are highly heterogeneous with respect to latent trait composition. A high c value yields a cluster with high $H_j s$, but as a consequence, many items sensitive to the same latent trait may be rejected. In general, when determining an appropriate value of c, a researcher should find a balance between the number of items in a scale and the strength of the scale.

Second, because MSP uses a sequential item selection procedure comparable to forward stepwise regression in SPSS (1998), not all combinations of items are considered. Therefore, the final item clusters may not have the maximum possible H coefficient for each cluster given all possible partitions of the total set. MSP offers a possibility to refine the search procedure; see Mokken (1971, pp. 198-199) and Sijtsma and Molenaar (2002, p. 72) for more details.

DETECT

Let composite θ_α be a linear combination of the separate θs from latent trait vector $\boldsymbol{\theta}$ (which may contain several dominant traits and several nuisance traits simultaneously). Composite θ_α can be understood as the latent direction that is best measured by the test (see Zhang & Stout, 1999a, for a rigorous definition of the direction of best measurement of a test). Given unidimensionality, following equation (2), the expected conditional covariance of an item pair equals 0. If θ_α is built up from multiple traits differentially measured by different items, the expected conditional covariance is positive when items j and k are driven by the same latent trait or traits that correlate highly, and it is negative when items j and k are driven by traits that correlate weakly or zero. The computer program

Figure 1
Geometrical Representation of Two Traits and Five Items



DETECT uses the sign behavior of the conditional covariances to find clusters of dimensionally homogeneous items.

More specifically, DETECT (Kim, 1994; Zhang, 1996; Zhang & Stout, 1999b) partitions, as much as possible, the set of items into an a priori specified maximum number of clusters in such a way that the expected conditional covariances between items from the same cluster are positive, and the expected conditional covariances between items from different clusters are negative. Consider an arbitrary partitioning $\mathcal{P}$ of the item pool. Let $\delta_{jk}(\mathcal{P}) = 1$ if items j and k are in the same cluster of $\mathcal{P}$, and $\delta_{jk}(\mathcal{P}) = -1$ otherwise (Zhang & Stout, 1999b). Then, the theoretical DETECT index is defined as

$$D_\alpha(\mathcal{P}) = \frac{2}{J(J-1)} \sum_{1 \le j < k \le J} \delta_{jk}(\mathcal{P}) E[\mathrm{Cov}\left(X_j, X_k | \theta_\alpha\right)]. \tag{6}$$

DETECT tries to find the partition that maximizes $D_\alpha(\mathcal{P})$. This partition is denoted as $\mathcal{P}^*$ and is taken as the final cluster solution. Thus, DETECT attempts to find dimensionally homogeneous clusters of items, each of which may be interpreted to assess another latent trait; this way, DETECT finds the number of dominant latent variables within a data matrix. Because the number of possible partitions increases very fast with the number of items, DETECT uses a genetic algorithm to search for the optimal partition. The criterion that is used to evaluate each partitioning is the DETECT index, $D_\alpha(\mathcal{P})$.

A geometrical representation (e.g., Ackerman, 1996; Stout et al., 1996), depicted in Figure 1, helps to visualize item response data driven by two θs. The vectors' length depends on the item discrimination, and the vectors' angles reflect the correlation between variables. Items j, k, l, m, and n are differentially sensitive to both θs, and item n exactly measures composite θ_α. The expected conditional covariance of an item pair is positive when the items are on the same side of θ_α (e.g., items j and k) and negative when the items are on different sides of θ_α. The expected conditional covariance of an item pair that includes item n equals zero. In practice, items that coincide with θ_α are rare (Zhang & Stout, 1999b).

Let rest score $R_{(-j,-k)} = X_+ - X_j - X_k$ be the total score ignoring the two studied items j and k. The sample DETECT statistic uses the following estimate of the expected conditional covariances,

$$E\left[\widehat{\mathrm{Cov}}(X_j, X_k | \theta_\alpha)\right] = \frac{E\left\{\widehat{\mathrm{Cov}}[X_j, X_k | R_{(-j,-k)}]\right\} + E\left[\widehat{\mathrm{Cov}}(X_j, X_k | X_+)\right]}{2}. \tag{7}$$

This average of the expected covariances was used because $E[\widehat{\text{Cov}}(X_j, X_k|X_+)]$ tends to be negatively biased and $E\{\widehat{\text{Cov}}[X_j, X_k|R_{(-j,-k)}]\}$ positively biased (Junker, 1993; Zhang & Stout, 1999a). The average of the two expected conditional covariances was expected to be less biased (Zhang & Stout, 1999a).

Additional remarks. First, DETECT is relatively new, and much theoretical research remains to be done. For example, the distribution of theoretical $D_\alpha(\mathcal{P})$ under interesting hypotheses is still unknown. In addition, in spite of equation (7), the DETECT index still is slightly biased (e.g., Zhang, Yu, & Nandakumar, 2003, investigate bias for various DETECT indices).

Second, Zhang and Stout (1999b) showed that DETECT finds the correct partitioning if items are mainly sensitive to one trait and only marginally to other traits. This is known as approximate simple structure (see Zhang & Stout, 1999b for a rigorous definition). When data deviate from approximate simple structure, the correct dimensionality may not be found (Zhang & Stout, 1999b).

Third, the DETECT index expresses the magnitude of the departure from unidimensionality within one or more clusters of the partition but is not an index of the number of traits within the item response data. Thus, there may be a high number of dimensions and yet $D_\alpha(\mathcal{P})$ is small, or there may be few dimensions and yet $D_\alpha(\mathcal{P})$ is large.

HCA/CCPROX

The software package HCA/CCPROX (Roussos et al., 1998) uses agglomerative hierarchical cluster analysis (HCA) for finding clusters of items. The program provides the opportunity to choose between different statistics, including conditional covariances, for assessing the relationship between variables. The user can also choose between different agglomerative HCA methods. Only the combination of statistic and method that, according to Roussos et al. (1998), was most successful in dimensionality assessment is presented here.

The program starts with each of the J items as a separate cluster. Then, at the second level of hierarchy, the two items having the smallest expected conditional covariance, $E\{\text{Cov}[X_j, X_k|R_{(-j,-k)}]\}$, are joined. For the subsequent steps, some additional notation is introduced. In general, at one particular step in the clustering process, let A_v and A_w denote two clusters of items, containing J_v and J_w items, respectively. Let $R_{(-A_v,-A_w)}$ denote the rest score, containing all responses to items that are not in A_v and A_w. Then, the expected conditional covariance may be defined: $E\{\text{Cov}[X_i, X_j|R_{(-A_v,-A_w)}]\}$. In each of the subsequent levels of hierarchy, that pair of clusters is joined that is closest of all pairs according to the proximity measure,

$$\text{Prox}(A_v, A_w) = (J_v J_w)^{-1} \sum_{i \epsilon A_v} \sum_{j \epsilon A_w} |E[\text{Cov}(X_i, X_j|R_{(-A_v,-A_w)})]|.$$

The process of joining clusters is repeated until all J items are collected into one large cluster.

Additional remarks. First, HCA/CCPROX does not provide a formal criterion, such as the lower bound c of coefficient H in MSP or the maximum DETECT index $D_\alpha(\mathcal{P}^*)$, that helps the researcher to decide which one of the $J - 1$ possible cluster outcomes reflects the true dimensionality best. Consequently, the researcher must choose the solution that most likely represents the dimensionality of the item response data. Due to the lack of a formal criterion, the researcher should rely on a priori theoretical expectations about the true dimensionality structure of the data. For example, when it is expected that a verbal test measures vocabulary, grammar, and spelling, and each item is assumed to predominantly measure one trait, then the three-cluster solution from HCA/CCPROX is appropriate here.

Second, according to Roussos et al. (1998), the positively biased $E\{\widehat{\text{Cov}}[X_i, X_j|\ R_{(-A_v,-A_w)}]\}$ will not affect the cluster analysis much because two items sensitive to different traits have an expected conditional covariance that is larger than that of two items that are sensitive to the same

latent trait. HCA/CCPROX should therefore be able to correctly partition the items according to their dimensionality.

DIMTEST

DIMTEST is a statistical test procedure that evaluates the unidimensionality of data from a user-specified item set (Nandakumar & Stout, 1993; Stout, 1987; Stout et al., 2001). The procedure of DIMTEST is the following. First, the item pool is split into three subtests, of which two are assessment subtests (denoted AT1 and AT2) and one is a partitioning subtest (denoted PT). One may use factor analysis or, for example, MSP or DETECT to have a sensible basis for AT1, AT2, and PT. DIMTEST provides linear factor analysis on the tetrachoric correlation matrix to determine which M items out of the total set of J items (the number M is user specified; for rules of thumb, see Nandakumar & Stout, 1993) are selected in AT1. These M items that constitute AT1 are hypothesized to be sensitive to the same trait. AT2 consists of M items sensitive to another trait than that measured by AT1 but with a similar observed frequency distribution of proportions correct on the items. Subtest PT is formed using the $J - 2M$ remaining items.

Using the sum scores on the PT subtest, the group of examinees is partitioned into subgroups of at least 20 (as recommended by Stout, 1987) of approximately equal ability. AT2 is designed to reduce "examinee variability bias" (i.e., θ still has a positive variance given a fixed PT score) and "item difficulty bias" (i.e., θ variance is inflated even more when items in the AT1 test and the PT test vary in difficulty). For short tests, both kinds of bias may inflate the DIMTEST statistic enough to incorrectly reject the null hypothesis of unidimensionality.

Let X_j^{AT1} and X_k^{AT1} be the scores on two items from AT1, and let Y_{PT} be a total score comparable with X_+ based on all items in PT. The DIMTEST sample statistic is based on

$$\mathrm{Cov}\left(X_j^{AT1}, X_k^{AT1} | Y_{PT} = y\right). \tag{8}$$

Under unidimensionality and for large J, this covariance must be close to zero for any item pair from AT1 and any Y_{PT} score. Under regularity conditions, the original DIMTEST statistic T (Stout, 1987) and the more powerful T' (Nandakumar & Stout, 1993) are distributed asymptotically (both in N and J) standard normally when unidimensionality holds. Given a significance level α and the upper $100(1 - \alpha)$ percentile of a standard normal distribution, Z_α, unidimensionality is rejected when $T > Z_\alpha$ or $T' > Z_\alpha$.

Additional remarks. First, DIMTEST tests the specific hypothesis that unidimensionality holds in a particular data set. For that reason, DIMTEST, unlike MSP, DETECT, and HCA/CCPROX, cannot directly be used to partition items in different clusters. Second, DIMTEST exhibits some positive bias because of the use of test scores as a conditioning variable even after correcting for two types of bias using AT2. Third, Stout et al. (2001) proposed a new DIMTEST procedure that uses only one subtest AT. The aim of the new DIMTEST procedure is to further reduce bias and increase power of T'. The properties of the new procedure are still subject to investigation. Therefore, the new procedure was not used in this study.

A Simulation Study

A simulation study was done to compare the effectiveness of MSP, DETECT, and HCA/CCPROX for selecting items into clusters that represent the true dimensionality of the data. Also, it was investigated whether the DETECT statistic, $D_\alpha(\mathcal{P})$, and the DIMTEST statistic, T', indicate whether the true model is essentially unidimensional or multidimensional. The simulation study involved six factors: (a) the IRT model used for simulating the data (two models), (b) the number of latent

traits (two numbers), (c) the correlation between the latent traits (six correlations), (d) the number of items per trait (for each number of latent traits, four combinations of numbers of items), (e) the item discrimination per trait (three combinations), and (f) the item selection method (four methods). For each cell of the $2 \times 2 \times 6 \times 4 \times 3 \times 4$ design, 2,000 simulees were generated from a multivariate standard normal density. Data were simulated assuming simple structure (Stout et al., 1996), meaning that items loaded only on one trait, but traits were allowed to correlate. Part of the design was replicated five times to investigate the stability of the results. For a few cells of the design, a smaller sample size ($n = 200$) was investigated.

IRT model. To simulate multidimensional item response data, the multidimensional extensions of the 2-PLM and the five-parameter acceleration model (5-PAM) (see also Sijtsma & Van der Ark, 2001; Samejima, 1995, 2000) were used. Several researchers (e.g., Hemker et al., 1995; Reckase & McKinley, 1991; Roussos et al., 1998) used the 2-PLM for simulating data, but data were also simulated using the more general 5-PAM to allow IRFs to take on a more flexible shape. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_Q)$ be the vector of Q latent traits (no nuisance traits), and let θ_{iq} be the value of person i on trait q. The 5-PAM has five item parameters: Let α_{jq} be the discrimination parameter of item j on trait q ($q = 1, \ldots, Q$); δ_{jq} the location parameter of item j on trait q; γ_j^{up} and γ_j^{lo} the upper and lower asymptotes of the IRF, respectively; and ξ_j the acceleration parameter. Then, for a multidimensional extension of the 5-PAM, to be denoted M5-PAM, the probability of answering item j correctly, given the latent trait vector $\boldsymbol{\theta}$, is

$$P(X_j = 1|\boldsymbol{\theta}) = \gamma_j^{lo} + (\gamma_j^{up} - \gamma_j^{lo}) \left\{ \frac{\exp\left[\sum_{q=1}^{Q} 1.7\alpha_{jq}(\theta_{iq} - \delta_{jq})\right]}{1 + \exp\left[\sum_{q=1}^{Q} 1.7\alpha_{jq}(\theta_{iq} - \delta_{jq})\right]} \right\}^{\xi_j}. \quad (9)$$

Parameter γ_j^{lo} and parameter γ_j^{up} allow the lower asymptote to be larger than 0 and the upper asymptote to be smaller than 1, respectively. Parameter ξ_j allows the IRF to be asymmetric (see also Samejima, 1995, 2000). The multidimensional 2-PLM (M2-PLM) (see also Reckase, 1997) is a special case of the M5-PAM for $\gamma_j^{lo} = 0$, $\gamma_j^{up} = 1$, and $\xi_j = 1$.

Number of traits. The numbers of latent traits used here were two and four.

Correlation between traits. The six product-moment correlations (ρ) between the latent traits were 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. The correlation of 0.0 represents independent latent traits, and the correlation of 1.0 represents unidimensionality.

Number of items per trait. For $Q = 2$ and $Q = 4$, four different combinations of the number of items per trait were chosen. Each trait was measured by either a small or a large number of items. For $Q = 2$, the four different combinations of test lengths within the item pool were as follows: short-short, short-long, long-short, and long-long. Notation [2:v; w] was used to indicate that two latent traits were generated with v items sensitive to θ_1 and w items sensitive to θ_2. Likewise, [4:v; w; y; z] is the four-dimensional extension of this notation. For $Q = 2$, the four combinations were [2:7;7], [2:7;21], [2:21;7], and [2:21;21]; for $Q = 4$, the four combinations were [4:7;7;7;7], [4:7;7;21;21], [4:21;21;7;7], and [4:21;21;21;21]. Each of these eight simulated combinations of the number of items per trait is referred to as the *true dimensional structure* or the *simulated dimensional structure*. It may be noted that by varying the number of items per trait across design cells, the total number of items in the item pool across design cells also varies.

Discrimination per trait. All items measuring the same latent trait either had low discrimination or high discrimination. If items all had low discrimination, the discrimination parameters were sampled from a distribution, to be discussed shortly, in such a way that discrimination varied but was low for all items. The same procedure was followed for items having high discrimination. Once the

parameters had been sampled, they were fixed across the design cells for which the discrimination level was held constant. Information referring to high-discrimination items is printed in boldface. For example, for $Q = 2$ and seven items per subset, three combinations of discrimination were used: [2:7;7], [2:7;**7**], and [2:**7**;**7**]; for $Q = 4$, the combinations were [4:7;7;7;7], [4:7;7;**7**;**7**], and [4:**7**;**7**;**7**;**7**].

Item discrimination was operationalized as the maximum slope of the IRF. In the special case of the M2-PLM, this maximum equals the discrimination parameter α_{jq}, but in the M5-PAM, the slope also depends on parameters γ_j^{lo}, γ_j^{up}, and ξ_j. Thus, in the M5-PAM, the maximum slope (α_{jq}^*) was calculated using the first partial derivative of equation (9). This resulted in

$$\begin{aligned}\alpha_{jq}^* &= \frac{4}{1.7}\left[\max\left(\frac{\partial P_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\right] \\ &= \frac{4}{1.7}\left[\alpha_{jq}\xi_j(\gamma_j^{up}-\gamma_j^{lo})\left(\frac{\xi_j}{1+\xi_j}\right)\left(1-\frac{\xi_j}{1+\xi_j}\right)\right]. \end{aligned} \tag{10}$$

From equation (10), it follows that

$$\alpha_{jq} = \frac{\alpha_{jq}^*}{\frac{4}{1.7}\left[\xi_j(\gamma_j^{up}-\gamma_j^{lo})(\frac{\xi_j}{1+\xi_j})(1-\frac{\xi_j}{1+\xi_j})\right]}. \tag{11}$$

Thus, α_{jq} can be calculated when γ_j^{lo}, γ_j^{up}, ξ_j, and α_{jq}^* are known. Constant $4/1.7$ is included in equation (11), so that in the M2-PLM, $\alpha_{jq}^* = 1.7 \times \alpha_{jq}$. Thus, α_{jq} depends on γ_j^{lo}, γ_j^{up}, ξ_j, and α_{jq}^*.

Parameters γ_j^{lo}, γ_j^{up}, and ξ_j influence the location of θ where α_{jq}^* reaches its maximum. If δ_{jq}^* is the location where the M5-PAM item discriminates best, then the corresponding location parameter equals

$$\delta_{jq} = \delta_{jq}^* - \frac{\ln(\xi_{jq})}{\alpha_{jq}^*}. \tag{12}$$

The parameters were generated to resemble parameter estimates found in the analysis of real test data. Under the M2-PLM, for items with low discrimination, α_{jq} is the exponentiation of a number randomly drawn from a normal distribution with mean 0.75 and variance 0.1, truncated at 0.5 and 1.25. For items with high discrimination, α_{jq} is the exponentiation from a number randomly drawn from a normal distribution with mean 1.75 and variance 0.1, truncated at 1.5 and 2.25. The difficulty parameters were chosen equidistant between -2.0 and 2.0.

Under the M5-PAM, γ_j^{lo} was chosen from the interval between 0.0 and 0.2, γ_j^{up} was chosen between 0.8 and 1.0, and ξ_j was chosen between 0.5 and 7, such that the slope (α_{jq}^*) and the location (δ_{jq}^*) under the M2-PLM and the M5-PAM were mathematically equal. However, the different shapes of the curves may prevent a direct and easy comparison of the results generated under the two models.

Item selection method. For the three item selection procedures (MSP, DETECT, and HCA/CCPROX) and for DIMTEST, the default settings were used as much as possible. Also, the recommendations made by the authors in various papers were used.

For MSP, the default lower bound value of $c = 0.30$ was used (Molenaar & Sijtsma, 2000). In addition, following recommendations by Hemker et al. (1995), for a part of the design, the influence of different c-values (0.10, 0.20, 0.30, 0.40, and 0.50) on the retrieval of the true dimensionality structure was investigated.

For DETECT, DIMTEST, and HCA/CCPROX, stable conditional covariance estimates were obtained using the item score vectors of at least 20 simulees per estimated θ_α (Stout, 1987) unless

this led to the rejection of more than 15% of the item score vectors. Then, the minimum group size was lowered to 10.

For DIMTEST, factor analysis of 500 item score vectors determined which items were used in AT1. The remaining 1,500 item score vectors were used to calculate the DIMTEST statistic. As recommended by Nandakumar and Stout (1993), the number of items M included in AT1 was determined by the rules that $4 < M \leq J/4$ and the absolute value of the loadings $\geq .15$. In the 14-item tests, $M = 3$ was used.

Results

Comparison of the Item Selection Methods

In the notation [4: v, w; y; z], the first number (here, 4) reflects the number of clusters found by MSP, DETECT, or HCA/CCPROX; v reflects the number of items selected into the first cluster; w reflects the number of items selected into the second cluster; and so on. A semicolon separates two clusters that are sensitive to different latent traits. A comma separates two clusters that are sensitive to the same latent trait. A classification error is defined as two items in the same cluster that are sensitive to different latent traits. Such errors are denoted by a slash, as in [2:7/7], meaning that at least one of the two clusters contains items that are sensitive to different θs.

Five types of results are distinguished. *Type A* means all J items were selected into the true dimensional structure. *Type B* indicates that the correct number of clusters and no classification errors were found, but not all J items were selected. *Type C* reflects that the true dimensionality was found to a high degree, but the number of clusters was larger than the Q latent traits in the sense that two or more clusters were driven by the same trait. Thus, Types A, B, and C do not have classification errors. *Type D* reflects that the true dimensional structure was not found; that is, items driven by different latent traits were selected into one subset. *Type E* represents the result whereby all items were selected into one subset. Types D and E have classification errors. For $\rho = 1.0$, Type E is the correct outcome, and for $\rho = 0.0$, Type A is the correct outcome.

Two-Dimensional Data Sets Based on M2-PLM

Correlation between traits. Table 1 shows that as correlations between traits (ρ) increased, the simulated dimensional structure was found less often by each of the item selection procedures.

Interaction of Correlation Between Traits × Method. The effect of increasing ρ on item selection was more apparent in MSP than in DETECT and HCA/CCPROX. For example, MSP found the simulated structure in [2:**7**;**7**] for $\rho = 0.0$ and $\rho = 0.2$, and as ρ increased, MSP tended to select more items sensitive to different traits into the same cluster until, for $\rho = 1$, a Type E result was found. These classification errors are made when the interitem correlations are such that lower bound c is not restrictive enough to split items sensitive to different traits into different clusters. DETECT and HCA/CCPROX found the simulated structure approximately until $\rho = 0.8$. Table 1 shows that for highly correlating traits, DETECT continued to form multiple clusters, even when items correlated $\rho = 1.0$. Due to sampling fluctuations and a weakly biased $D_\alpha(\mathcal{P})$ statistic, the observable conditional covariances were nonzero, even when the data were unidimensional. For these reasons, $D_\alpha(\mathcal{P})$ can be highest for a partitioning having two or more clusters.

Discrimination. With increasing α^*_{jq}, the simulated dimensional structure was found more often for each of the item selection methods; see Table 1.

Interaction of Discrimination × Method. MSP was more sensitive to item discrimination than DETECT and HCA/CCPROX. Variation in mean α^* between latent traits within one data matrix

Table 1

Item Selection Results Using the Multidimensional Two-Parameter Logistic Model (M2-PLM) and Two Latent Traits

	ρ					
Test Composition	0.0	0.2	0.4	0.6	0.8	1.0
MSP						
[2:7; 7]	[3:2,5;6]	[3:2,5;7]	[2:7;6]	[3:2/3/7]	[4:2/2/2/8]	[2:10/2]
[2:7; 21]	[2:6;19]	[4:2,5;2,19]	[5:2,5;2,17]	[4:2/2/3/20]	[4:2/2/2/21]	[3:2/2/24]
[2:21; 7]	[3:19,2;7]	[3:19,2;5]	[2:20/5]	[3:20/4/2]	[3:22/2/2]	[2:25/2]
[2:21; 21]	[4:2,18;2,19]	[3:2,18;19]	[4:2,18; 2,19]	[4:2/2/9/27]	[5:2/2/2/2/31]	[2:2/39]
[2:7; **7**]	[3:2,5;7]	[2:7;6]	[2:6;7]	[1:13]	[1:14]	[1:14]
[2:7; **21**]	[2:6;21]	[2:7;21]	[2:5;21]	[2:2/25]	[1:27]	[1:28]
[2:21; **7**]	[3:2,18;7]	[3:2,18;7]	[4:2,2,17;7]	[2:2/26]	[1:27]	[1:27]
[2:21; **21**]	[3:2,19;21]	[3:2,18;21]	[3:2/17/23]	[3:2/2/37]	[2:2/40]	[1:42]
[2:**7**; **7**]	[2:7;7]	[2:7;7]	[1:14]	[1:14]	[1:14]	[1:14]
[2:**7**; **21**]	[2:7;21]	[2:7;21]	[1:28]	[1:28]	[1:28]	[1:28]
[2:**21**; **7**]	[2:21;7]	[2:21;7]	[1:28]	[1:28]	[1:28]	[1:28]
[2:**21**; **21**]	[2:21;21]	[2:21;21]	[1:42]	[1:42]	[1:42]	[1:42]
DETECT						
[2:7; 7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[3:3/5/6]	[5:2/2/2/2/6]
[2:7; 21]	[2:7;21]	[2:7;21]	[3:7;1,20]	[2:7;21]	[4:7;1,6,14]	[4:4/5/6/13]
[2:21; 7]	[2:21;7]	[2:21;7]	[2:21;7]	[3:2,19;7]	[4:2,2,19;7]	[4:3/10/10/5]
[2:21; 21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[4:1/12/12/17]
[2:7; **7**]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[3:1,6;7]	[3:2/3/9]
[2:7; **21**]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[4:2,2,3;21]
[2:21; **7**]	[2:21;7]	[2:21;7]	[2:21;7]	[4:1,2,18;7]	[4:4,8,9;7]	[4:3/3/4/18]
[2:21; **21**]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[3:5/8/29]
[2:**7**; **7**]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[1:14]
[2:**7**; **21**]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[3:3/11/14]
[2:**21**; **7**]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:10;18]
[2:**21**; **21**]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[3:18/16/8]
HCA/CCPROX						
[2:7; 7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:1/13]	[2:2/12]
[2:7; 21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:3/25]
[2:21; 7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:4/24]
[2:21; 21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:2/40]
[2:7; **7**]	[2:7,7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:2/12]
[2:7; **21**]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:6/22]
[2:21; **7**]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:4/24]
[2:21; **21**]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:9/32]
[2:**7**; **7**]	[2:7;7]	[2:7;7]	[2:7;7]	[2:7;7]	[2:2/12]	[2:2/12]
[2:**7**; **21**]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:7;21]	[2:10/18]
[2:**21**; **7**]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:21;7]	[2:3/25]
[2:**21**; **21**]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:21;21]	[2:5/37]

Note. Boldface indicates highly discriminating items. Bracket notation: a *semicolon* separates dimensionally different clusters, a *comma* separates dimensionally similar clusters, and a *slash* separates mixed clusters.

was also simulated. Latent traits that were represented by clusters of weakly discriminating items were not well recovered by any of the three item selection methods, but latent traits that were represented by means of highly discriminating items were well recovered.

Number of items per trait. Traits represented by 7 items were, in general, equally well recovered as traits represented by 21 items.

Interaction of Number of Items Per Trait $\times$ *Method*. For clusters containing 21 items having low item discrimination, MSP sometimes misclassified a single item out of the total set. Another result was that MSP selected the lowly discriminating items into an extra cluster (i.e., Type C). Such results were not found for latent traits assessed by 7 items. DETECT produced more Type C results in the unequal number of items condition compared to the equal conditions. HCA/CCPROX produced approximately the same results, irrespective of the number of items per trait.

Method. In general, the simulated structure was found more often by DETECT and HCA/CCPROX than by MSP. HCA/CCPROX results should be interpreted with care because only the outcomes were presented when the number of clusters equalled the number of simulated traits (Q). In practical data analysis, however, the researcher has to decide which cluster solution is best, possibly relying on previous knowledge about the trait structure of the data. Thus, the results of HCA/CCPROX presented here and elsewhere in the Results section may be more favorable than in practical data analysis. For $\rho = 1.0$, the HCA/CCPROX partitioning only reflects random fluctuation.

Replications Based on M2-PLM

For [2:7;7], [2:7;21], and [2:**7**;**7**]; for $\rho = 0.0, 0.4$, and 0.8; and for MSP, DETECT, and HCA/CCPROX, five data matrices were randomly and independently sampled (results are not presented in a table). True dimensionality was found consistently across replications, particularly for highly discriminating items and low correlations between traits. DETECT and HCA/CCPROX yielded more consistent results than MSP. This may be due to the scaling condition $H_j \geq c$ in MSP. For some items, this condition may be satisfied in some samples but not in others, resulting in different cluster solutions between samples. DETECT and HCA/CCPROX do not have such a scaling condition, and the effect of sample fluctuations on the cluster solution may therefore be smaller. In other design cells also included in the replication investigation, MSP and DETECT often found an extra cluster, and HCA/CCPROX misclassified several items.

Small Sample Size

The MSP and HCA/CCPROX results for $n = 200$ and $n = 2,000$ were approximately the same in the design cells for [2:7;7], [2:7;21], and [2:**7**;**7**], and $\rho = 0.0, 0.4$, and 0.8. DETECT's results were somewhat worse for $n = 200$, probably due to inaccurate conditional covariance estimates in too small X_+ and $R_{(-j,-k)}$ score groups. MSP uses the H_j coefficient, which is based on the whole sample and, therefore, is more stable.

Four-Dimensional Simulation Using the M2-PLM

In general, the results for $Q = 2$ and $Q = 4$ (see Table 2) were comparable. However, for $Q = 4$, more results of Type B and Type C were found (A, B, C, D, E notation is used to save space) because the greater number of items gave rise to more chance capitalization. A peculiar result for DETECT was that for [4:7;7;21;21] and [4:21;21;7;7]; as ρ increased, DETECT (but not HCA/CCPROX) selected the two clusters of seven equally discriminating items, sensitive to different latent traits with equal discrimination, into one cluster. The effect was more pronounced for higher discrimination. For HCA/CCPROX, only the correct (Type A) or incorrect (Type D) solutions were reported because of the use of the foreknowledge that $Q = 4$.

Table 2

Four-Dimensional Item Selection Results Using the Multidimensional Two-Parameter Logistic Model (M2-PLM)

Test Composition	MSP (ρ)						DETECT (ρ)						HCA/CCPROX (ρ)					
	.0	.2	.4	.6	.8	1	.0	.2	.4	.6	.8	1	.0	.2	.4	.6	.8	1
[4:7; 7; 7; 7]	B	C	B	D	D	D	A	A	A	A	A	D	A	A	A	A	D	D
[4:7; 7; 21; 21]	C	C	C	D	D	D	D	D	A	D	D	D	A	A	A	A	D	D
[4:21; 21; 7; 7]	C	C	C	D	D	D	A	D	D	A	A	D	A	A	A	A	A	D
[4:21; 21; 21; 21]	C	C	D	D	D	D	A	A	D	A	A	D	A	A	A	A	D	D
[4:7; 7; **7**; **7**]	C	C	D	D	D	E	A	A	A	A	D	D	A	A	A	A	D	D
[4:7; 7; **21**; **21**]	B	C	C	D	E	E	A	D	D	D	D	D	A	A	A	D	D	D
[4:21; 21; **7**; **7**]	C	C	D	D	D	D	A	D	D	A	A	D	A	A	A	A	D	D
[4:21; 21; **21**; **21**]	C	B	D	D	D	E	A	A	A	A	A	D	A	A	A	A	A	D
[4:**7**; **7**; **7**; **7**]	A	A	D	E	E	E	A	A	A	A	A	D	A	A	A	A	A	D
[4:**7**; **7**; **21**; **21**]	A	A	D	E	E	E	A	A	D	A	A	D	A	A	A	D	D	D
[4:**21**; **21**; **7**; **7**]	A	A	E	E	E	E	A	D	D	D	D	D	A	A	A	A	D	D
[4:**21**; **21**; **21**; **21**]	A	A	E	E	E	E	A	D	A	A	A	D	A	A	A	D	A	D

Note. Boldface indicates highly discriminating items; A = true dimensionality found; B = not all items included; C = multiple clusters; D = dimensionality not found; E = all items in one subset.

Two-Dimensional Simulation Using the M5-PAM

For data generation using the M5-PAM, only those factor levels were used that proved to be informative in the M2-PLM analysis: two traits (not four), either low or high discrimination (maximum slope α^*) (no combination), 7 or 21 items per trait, and correlations between the traits that varied from 0.0 to 1.0. The design, therefore, had the order 2 (discrimination levels) $\times$ 2 (number of items per trait) $\times$ 6 (correlation between traits) $\times$ 4 (item selection method).

The general trend in the results (see Table 3) was the same as with the simulation using the M2-PLM. For any of the three methods, for a higher ρ and a lower α^*, the dimensional structure was found less often (see Table 3). As before, these trends were more obvious for MSP than for DETECT and HCA/CCPROX. For the number of items per cluster, the effects were reversed: For 21-item clusters, somewhat better results were obtained than for 7-item clusters. However, the differences were small and may be due to sample fluctuation. As for the M2-PLM, DETECT found the simulated dimensionality less often for unequal numbers of items.

Compared to the M2-PLM, in general, all three methods performed a little worse. For MSP, more Type B results were found; for DETECT, more Type C results were found; and for HCA/CCPROX, more Type D results were found (cf. Tables 1 and 3). These results may, in part, be due to the different overall shapes of the IRFs of the M5-PAM and the M2-PLM. Even when two IRFs from different models have equivalent maximum slopes (and equal locations), their slopes are not the same for all θs. In this study, this resulted in a somewhat lower overall discrimination for the M5-PAM items. This might explain that more minor deviations from the simulated dimensional structure were found when using the M5-PAM than when using the M2-PLM.

Table 3

Two-Dimensional Item Selection Results Using the Multidimensional Five-Parameter Acceleration Model (M5-PAM)

Test Composition	MSP (ρ)						DETECT (ρ)						HCA/CCPROX (ρ)					
	.0	.2	.4	.6	.8	1	.0	.2	.4	.6	.8	1	.0	.2	.4	.6	.8	1
[2:7; 7]	B	B	B	E	E	E	A	A	A	C	C	D	A	A	A	A	D	D
[2:7; 21]	B	B	B	E	E	E	A	C	C	C	C	D	A	A	A	A	D	D
[2:21; 7]	B	B	E	E	E	E	C	C	C	C	D	D	A	A	A	A	A	D
[2:21; 21]	B	C	B	E	E	E	A	A	A	A	A	D	A	A	A	A	A	D
[2:**7**; **7**]	A	B	E	E	E	E	A	A	A	A	C	D	A	A	A	A	D	D
[2:**7**; **21**]	B	B	D	E	E	E	A	A	C	C	C	D	A	A	A	A	D	D
[2:**21**; **7**]	A	B	E	E	E	E	A	A	A	C	D	D	A	A	A	A	A	D
[2:**21**; **21**]	B	A	D	E	E	E	A	A	A	A	A	D	A	A	A	A	A	D

Note. Boldface indicates highly discriminating items; A = true dimensionality found; B = not all items included; C = multiple clusters; D = dimensionality not found; E = all items in one subset.

Manipulating Lower Bound c in Mokken Scale Analysis

The effect of using different c-values (0.00, 0.20, 0.30, 0.40, and 0.50) (Hemker et al., 1995) on MSP item selection was investigated for several design cells; that is, item discrimination was either low or high, seven items per trait were used, and the M2-PLM was used for generating data. This constituted a 3 (levels of mean discrimination) $\times$ 6 (correlation between traits) $\times$ 5 (lower bound c) design. In general, Table 4 shows that for low c-values, a high mean discrimination, and a high correlation between traits, all items were selected into one cluster (outcome Type E). As c increased, some items, depending on their discrimination and on the correlation between the latent traits, did not satisfy $H_j \geq c$ and, consequently, were not selected into this subset. When mean item discrimination was low and the correlation between latent traits was also low, with increasing c, many lowly discriminating items did not satisfy $H_j \geq c$ for the first subset but satisfied $H_j \geq c$ for the second subset. As a consequence, more clusters were found than simulated (outcome Type C). When item discrimination was high and lower bound c was also high, only the items sensitive to the same trait were collected into the same subset, thereby finding the true dimensional structure (outcome Type A). Thus, following Hemker et al. (1995), it was found that the choice of c greatly influences item selection results.

Comparing the DETECT and DIMTEST Test Statistics

DETECT. For data generated using the M2-PLM, Table 5 shows the values of $D_\alpha(\mathcal{P}^*)$ multiplied by 100; for simplicity, the result is also called $D_\alpha(\mathcal{P}^*)$. Following Zhang and Stout (1999a), $D_\alpha(\mathcal{P}^*) < 0.1$ is interpreted as essential unidimensionality and $D_\alpha(\mathcal{P}^*) > 1.0$ as sizable multidimensionality. Based on Douglas, Kim, Roussos, Stout, and Zhang (1999), $0.1 < D_\alpha(\mathcal{P}^*) < 1$ can be interpreted as moderate multidimensionality. Thus, to correctly interpret DETECT's results, the clustering solution as well as the value of $D_\alpha(\mathcal{P}^*)$ should be considered.

Table 4

MSP Results for Different Lower Bounds *c*

c	Test Composition	ρ = *0.0*	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*
0.10	[2:7; 7]	A	D	E	E	E	E
	[2:7; **21**]	A	D	E	E	E	E
	[2:**21**; **21**]	A	E	E	E	E	E
0.20	[2:7; 7]	A	A	D	E	E	E
	[2:7; **21**]	A	A	E	E	E	E
	[2 : **21**; **21**]	A	A	E	E	E	E
0.30	[2:7; 7]	C	C	B	D	D	D
	[2:7; **21**]	C	B	B	E	E	E
	[2:**21**; **21**]	A	A	E	E	E	E
0.40	[2:7; 7]	C	C	C	C	C	D
	[2:7; **21**]	B	C	C	D	D	E
	[2:**21**; **21**]	A	A	A	C	E	E
0.50	[2:7; 7]	B	C	C	C	B	B
	[2:7; **21**]	B	B	B	B	B	E
	[2:**21**; **21**]	A	A	A	C	E	E

Note. Boldface indicates highly discriminating items;
A = true dimensionality found; B = not all items included;
C = multiple clusters; D = dimensionality not found;
E = all items in one subset.

Table 5 shows that $D_\alpha(\mathcal{P}^*)$ is smaller as the correlation between latent traits is closer to 1.0. Also, $D_\alpha(\mathcal{P}^*)$ tends to be higher for high discrimination items and lower when clusters contained different numbers of items (i.e., [2:7;21]). Based on the rules of thumb, for equal numbers of items per trait, for $\rho = 0.0$, 0.2, and 0.4, statistic $D_\alpha(\mathcal{P}^*)$ correctly indicated sizable multidimensionality; for $\rho = 0.6$ and 0.8, statistic $D_\alpha(\mathcal{P}^*)$ indicated moderate multidimensionality; and for $\rho = 1.0$, statistic $D_\alpha(\mathcal{P}^*)$ often indicated unidimensionality. For unequal numbers of items, statistic $D_\alpha(\mathcal{P}^*)$ indicated moderate multidimensionality, except for $\rho = 1.0$, for which $D_\alpha(\mathcal{P}^*)$ indicated unidimensionality.

Because the values of $D_\alpha(\mathcal{P}^*)$ for $\rho = 1.0$ indicate unidimensionality, the clusters DETECT yielded for $\rho = 1.0$ and that were presented in Table 1 should be ignored. Unidimensionality was supported only for $\rho = 1.0$ when items discriminated highly or when clusters contained 21 items (see Table 1). The results in Table 5 show that for unidimensional data, $D_\alpha(\mathcal{P}^*)$ values were found as high as 0.202. This may indicate that the upper bound of 0.1 for essential unidimensionality may be too low. As one of the reviewers suggested, the rules of thumb for interpreting $D_\alpha(\mathcal{P}^*)$ may be the topic of future research.

DETECT works less well for unequal numbers of items. This result can be explained using Figure 2. Let [2:7;21] be the simulated dimensionality, then the direction of best measurement of the test, θ_α, lies closer to the direction of θ_2 because there are more items sensitive to this trait. As a consequence, the expected conditional covariances for items sensitive to θ_2 are closer to 0 than

Table 5

Results of DETECT Statistic $D_\alpha(\mathcal{P}^*)$ Using the Multidimensional Two-Parameter Logistic Model (M2-PLM) for Two Latent Traits

Test Composition	ρ					
	0.0	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*
[2:7; 7]	1.564	1.254	1.010	0.642	0.242	0.186
[2:7; 21]	0.756	0.654	0.535	0.379	0.203	0.108
[2:21; 7]	0.827	0.681	0.551	0.380	0.203	0.115
[2:21; 21]	1.889	1.513	1.083	0.731	0.361	0.092
[2:7; **7**]	1.836	1.547	1.111	0.775	0.441	0.202
[2:7; **21**]	0.905	0.806	0.621	0.427	0.303	0.118
[2:21; **7**]	0.851	0.723	0.574	0.361	0.208	0.119
[2:21; **21**]	2.132	1.702	1.341	0.879	0.442	0.090
[2:**7**; **7**]	2.137	1.750	1.266	0.842	0.385	0.067
[2:**7**; **21**]	0.978	0.826	0.668	0.472	0.259	0.039
[2:**21**; **7**]	0.999	0.816	0.653	0.494	0.247	0.042
[2:**21**; **21**]	2.400	2.016	1.508	0.967	0.518	0.034

Note. Boldface indicates highly discriminating items.

the expected conditional covariances for items sensitive to θ_1, and because of that, their dispersion also is smaller. Furthermore, item pairs from the larger clusters more easily may have negative conditional covariances, even though they are sensitive to the same latent trait, and items from different clusters may have positive expected conditional covariances, even though they are sensitive to different traits. Such incorrect sign behavior is more likely for item pairs that are sensitive to θ_2 because, due to their smaller angles with θ_α, they more often are on different sides of θ_α than item pairs that are sensitive to θ_1.

DIMTEST. Using DIMTEST statistic T', Table 6 shows that, in general, for $\rho = 1.0$, unidimensionality was found, and for $\rho \leq .8$, multidimensionality was found. These results were more pronounced for highly discriminating items. These results were not found for 14 items ([2:7;7]), maybe because T' is an asymptotic statistic.

Unidimensionality was found unexpectedly for [2:21;21] and [2:**21**;**21**] when $\rho = 0.4$. Given that AT1 must be as dimensionally distinct as possible from AT2 and PT, then both setups result in an AT1 that is sensitive to either θ_1 or θ_2 and an AT2 and a PT that are sensitive to a mixture of θ_1 and θ_2. This may result in less power to reject the null hypothesis in these cases (see also Nandakumar & Stout, 1993) because the covariance between AT1 items will be relatively low when PT scores are partly driven by the same latent trait. One may note that DIMTEST performs well when latent traits are represented with an unequal number of items and, to a lesser extent, with unequal discrimination. In the latter cases, AT1 and PT are, in a large degree, driven by distinct latent traits.

In Table 7, the number of times the null hypothesis was rejected using a nominal significance level of .05 is reported for five replicated data matrices. The results agree to a high degree with the findings presented in Table 6. For example, the results for equal numbers of items per trait are

Figure 2
Geometrical Representation of One Short Test (θ_1) and One Long Test (θ_2)

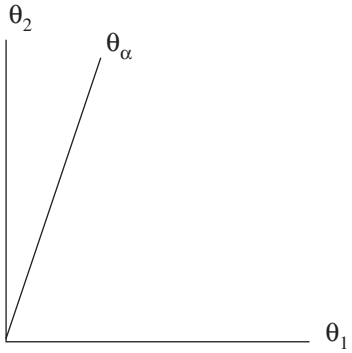


Table 6
Results of DIMTEST Statistic T' Using the Multidimensional Two-Parameter Logistic Model (M2-PLM) for Two Latent Traits

Test Composition	ρ					
	0.0	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*
[2:7; 7]	○	○	○	○	○	○
[2:7; 21]	●	●	●	●	○	○
[2:21; 7]	●	●	●	●	○	○
[2:21; 21]	○	●	○	●	●	○
[2:7; **7**]	●	○	●	●	○	○
[2:7; **21**]	●	●	●	○	○	○
[2:21; **7**]	●	●	●	●	●	○
[2:21; **21**]	●	●	●	○	○	○
[2:**7**; **7**]	○	●	○	○	○	○
[2:**7**; **21**]	●	●	●	●	●	○
[2:**21**; **7**]	●	●	●	●	●	○
[2:**21**; **21**]	●	●	○	●	●	○

Note. Boldface indicates highly discriminating items; ● denotes significant result using .05 as significance level (i.e., multidimensionality), ○ indicates a nonsignificant result (i.e., unidimensionality).

less stable than for unequal numbers of items per trait. Also, for [2:7;7], the results mainly reflect random fluctuation.

Table 7

Frequency (Out of Five) With Which DIMTEST Statistic T' Rejects the Null Hypothesis (5% Level), Using the Multidimensional Two-Parameter Logistic Model (M2-PLM) for Simulation

Test Composition	ρ					
	0.0	*0.2*	*0.4*	*0.6*	*0.8*	*1.0*
[2:7; 7]	2	4	0	1	0	0
[2:7; 21]	5	5	5	4	2	0
[2:21; 7]	5	5	5	5	2	0
[2:21; 21]	5	2	4	4	3	1
[2:7; **7**]	0	3	1	0	1	0
[2:7; **21**]	5	5	5	4	2	0
[2:21; **7**]	5	5	5	5	3	1
[2:21; **21**]	4	5	5	5	3	1
[2:**7**; **7**]	2	2	2	2	0	0
[2:**7**; **21**]	5	5	5	5	5	0
[2:**21**; **7**]	5	5	5	5	5	0
[2:**21**; **21**]	5	5	5	4	4	0

Note. Boldface indicates highly discriminating items.

Conclusion and Discussion

Using the methods as recommended in the literature, DETECT and HCA/CCPROX were superior to MSP in retrieving the simulated dimensional structure. Even when there was little information available to distinguish items that are sensitive to different traits (e.g., highly discriminating items, sensitive to two traits that correlated 0.8), DETECT and HCA/CCPROX retrieved the dimensional structure, but MSP failed. It may be noted, however, that traits correlating 0.8 may be indistinguishable from a substantive viewpoint, which puts the MSP result in a more positive perspective. In general, DETECT performed better than HCA/CCPROX, but in some instances (e.g., when discrimination was low and tests were long), HCA/CCPROX was superior to DETECT.

Differences between DETECT and HCA/CCPROX may be due to different conditional covariance estimates used in these methods. Also, DETECT's algorithm is less susceptible to locally optimal solutions than HCA/CCPROX's algorithm. In addition, in practice, the researcher must choose the final HCA/CCPROX cluster solution among $J - 1$ solutions. This may be an extra source for differences.

MSP and DETECT differ in many ways. First, MSP uses normed unconditional covariances, and DETECT uses conditional covariances. Second, MSP uses a sequential clustering procedure based on several item selection criteria, and DETECT searches for the item partition that maximizes $D_\alpha(\mathcal{P})$ without additional selection criteria. Third, MSP selects items that satisfy $H_j \geq c$, where c is meant as a minimum quality criterion for item discrimination. As a result, the default setting $c = 0.3$ may not, for example, select all items driven by the same trait in a cluster and thus may not yield the "true" dimensionality. Other, nondefault values of c, however, may yield the correct dimensionality. Fourth, items selected by MSP into a cluster cannot leave this cluster, and this

makes MSP susceptible to locally optimal solutions. DETECT uses a genetic algorithm, which moves items back and forth until a final solution is found. These differences make the comparison of MSP and DETECT difficult.

Two remarks with respect to the DETECT and DIMTEST statistics are in order. First, it was found that the number of items in a cluster assessing one trait may influence the assessment of dimensionality; DETECT's maximum, $D_\alpha(\mathcal{P}^*)$, did not reflect the dimensionality well for data matrices that contained clusters with an unequal number of items, and DIMTEST's T' did not reflect the dimensionality well for clusters with equal numbers of items and equal average discrimination. Second, the results for $n = 200$ made clear that DETECT and DIMTEST may be more effective in larger samples (here, $n = 2,000$ was investigated).

Practical recommendations. Based on the simulation study, DETECT and MSP were found to be the most useful programs for finding unidimensional item clusters. Both methods yield a single clustering solution and provide test statistics for evaluating the quality of the cluster solution. In general, DETECT recovered the simulated dimensionality better than MSP, but DETECT needed larger samples than MSP. Furthermore, for data sets with highly correlating latent traits, DETECT forces items into clusters, and therefore seems to be vulnerable to chance capitalization. MSP always produces the best item clustering according to the definition of a Mokken scale and discards items not fitting well. The methods were compared as they are available to researchers. Future research may use in one selection procedure conditional covariances to find the true dimensionality of the data and a minimum item quality criterion for only selecting practically useful items.

DIMTEST is suitable when the researcher expects his or her data to be unidimensional but cannot be used to partition the data. DIMTEST has low power for short tests. HCA/CCPROX yields $J - 1$ cluster solutions. This forces the researcher to make a choice about the dimensionality of the test. A drawback of the method is that it does not provide the value of a quality statistic for each solution, such as $D_\alpha(\mathcal{P}^*)$ or H, on which the researcher can base his or her choice.

At the practical level, it may be noted that MSP uses a Windows interface and can be run under Windows 95 or higher. DETECT, HCA/CCPROX, and DIMTEST are DOS programs.

Because the methods are so different, a practical recommendation is to use them next to one another to analyze the same data set. For example, one could use DETECT to find dimensionally distinct clusters and use MSP to select the best discriminating items for which $H_j > c$ within these clusters. HCA/CCPROX can be informative about the process of clustering (e.g., which items have most in common and are clustered first and which are added later). DIMTEST can be used to verify unidimensionality of the clusters, especially because this method has more power than DETECT when only few items are driven by another trait than the main trait.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory. *Applied Psychological Measurement*, *20*, 311-329.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

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.

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.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69-95). New York: Springer.

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

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*, 1-14.

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.

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

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear mutlidimensional IRT models for polytomously scored items. *Psychometrika*, *59*, 149-176.

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

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457-477.

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

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

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

Miecskowski, T. A., Sweeney, J. A., Haas, G., Junker, B. W., Brown, R. P., & Mann, J. (1993). Factor composition of the suicide intent scale. *Suicide and Life Threatening Behavior*, *23*, 37-45.

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

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

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

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, *48*, 49-72.

Molenaar, I. W., & Sijtsma, K. (2000). *User's 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.

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.

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

Roussos, L. A. (1992). *Hierarchical agglomerative clustering computer program user's manual*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

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. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, *60*, 549-572.

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

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

Sijtsma, K., & Van der Ark, L. A. (2001). In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 297-318). New York: Springer.

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 293-325.

Stout, W. F. (1990). A new item response theory modeling approach with applications to

unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST manual*. Unpublished manuscript, University of Illinois, Urbana-Champaign.

Stout, W. F., 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.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331-354.

Zhang, J. (1996). *Some fundamental issues in item response theory with applications*. Unpublished doctoral dissertation, University of Illinois, Urbana- Champaign.

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.

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.

Acknowledgments

The authors would like to thank Bas T. Hemker, Ivo W. Molenaar and William F. Stout, and three anonymous reviewers for their useful comments. This research was supported by a travel grant of the Dutch Organization for Scientific Research (NWO), grant number SIR 12-3988.

Author's Address

Address correspondence to A. A. H. van Abswoude, Department of Methodology and Statistics, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: A.A.H.vanAbswoude@uvt.nl.