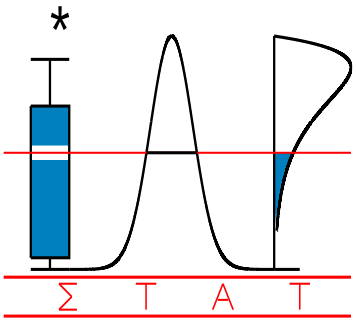
T E C H N I C A L

R E P O R T

0684

DETECTING LOCAL ITEM DEPENDENCE STEMMING FOR MINOR DIMENSIONS

BALAZS, K. and P. DE BOECK



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Running head: DETECTING LOCAL ITEM DEPENDENCE

Detecting local item dependence stemming from minor dimensions

Katalin Balázs and Paul De Boeck

K.U. Leuven, Belgium

Abstract

Local item dependence (LID) is a common violation of the local stochastic independence assumption for statistical models such as in item response models. LID may be induced by various factors and it can hardly be avoided in the practice of testing. Ignoring LID can have serious consequences for the goodness of fit of a model, for the parameter estimates and for confidence intervals. Suitable diagnostic methods are necessary to detect LID, so that appropriate action can be taken, such as an adapted model formulation.

In a simulation study, several potential LID diagnostics are investigated on their performance. The focus is on LID stemming from minor dimensions additional to the common dimension in the data. A group of items may constitute an additional dimension, for example because they are all related to the same stimulus or to the same but rather specific underlying processes. The purpose of the present study is to compare several diagnostic methods for a kind of data that is rather common in psychology, with a small sample size and a rather small set of items, and for a model that is in correspondence with the use of sum scores (the Rasch model). The methods that will be investigated fall into three categories: pairwise diagnostics, clustering techniques and visual methods.

From the pairwise diagnostics, the modified Mantel-Haenszel test seem to perform the best as an absolute detection criterion, and Yen's Q_3 and the standardized log-odds ratio difference (τ) seem to perform the best as a relative criterion. Among the clustering techniques, the DETECT procedure was not successful in revealing the studied type of LID, while ADCLUS had a reasonable performance, which was at its best when applied to the standardized log-

odds ratio difference (τ) or to the log-odds ratio. Finally, also the grey-scale matrices based on the pairwise statistics seem to be effective.

Local item dependence (LID) is a term to denote the remaining dependence in the data beyond the dependencies already explained by the model, often one with a general dimension. LID is a quite common violation of the assumptions of item response models and it can occur in many ways in practice. Because LID leads to unwanted effects, as it will be described later, it is important to have methods which can identify the problem. The aim of the present study is to investigate and compare a set of LID diagnostics for their performance in detecting a common kind of dependency in binary person by item data, and to study the data features which may affect their performance. First, a short general introduction will be provided to the theory of item dependencies, and afterwards a simulation study will be reported.

Local stochastic independence (LSI) is a basic assumption in item response theory, implying that the probability of an item response conditional on the general ability parameter is not influenced by other item responses (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000; Jiao & Kamata, 2003; Lazarfeld & Henry, 1968; Lord, 1980, Lord & Novick, 1968, Wilson & Adams, 1995). In other words, given the modeled relation between two items, no further relation may exist for LSI to apply. LSI is also termed conditional independence because the independence applies conditionally on the parameters (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968). When the LSI assumption is violated, there are remaining *local item dependencies* (LIDs).

Several different testing situations may lead to LID. Yen (1993) differentiates among numerous causes: external assistance, speededness, fatigue, practice, item or response format, passage dependence, items that require an explanation of the previous response, scoring rubrics, raters, content knowledge and abilities. In general, two main sources can be discerned. First, the dependency may stem from the items being located in an *item chain* where the success on an item can depend on the successes on previous items (Thissen et al, 1992) or, more in general, on their position in the chain. Second, items may show *item*

overlap in terms of their parts (e.g., a common item stem) or in terms of the processes underlying the item response. Hoskens and De Boeck (1997) have used the terms *order interaction* and *combination interaction* for these two types (item chain and item overlap), respectively.

As for the interpretation of LID, Ip, Wang, De Boeck and Meulders (2004) differentiate between LID as a side-effect of the design of items and tests, on the one hand, and more basic, substantive LID based on psychological processes, such as cognitive components, as discussed in Jannarone (1986) and Hoskens and De Boeck (1995, 2001), on the other hand. Chen & Thissen (1997) classified LID into surface local dependence (SLD) and underlying local dependence (ULD). SLD simply means that the examinees respond to a set of items very similarly, while ULD assumes a separate trait belonging to each locally dependent item group.

Zensky, Hambleton and Sireci (1991) pointed out that the problem of LID was already recognized in classical test theory: theorists knew that the estimates are not correct when item dependencies were not properly accounted for. Kelley (1924), Guilford (1936), Thorndike (1951) and Anastasi (1961) warned that items corresponding to a common stimulus should be placed in the same half of a test when computing split-half reliabilities.

IRT models are not robust to the violation of local item independence. In case of LID, the parameter estimations may be biased and the reliability of the estimates may be overestimated (or underestimated) (Chen & Thissen, 1997; Tuerlinckx & De Boeck, 1998; Yen, 1993). LID may yield biased parameter estimation and deteriorates equating (Ackerman, 1987; Chen & Thissen, 1997; Fennessy, 1995; Sireci, Thissen & Wainer, 1991; Spray & Ackerman, 1987; Thissen, Steinberg & Mooney, 1989; Tuerlinckx & De Boeck, 1998; Yen, 1984). Item banking and item calibrations also become difficult when LID occurs (Ferrara,

Huynh, & Baghi, 1997). In case of computerized adaptive testing, the biased standard error estimates may lead to premature termination (Fennessy, 1995).

Several approaches have been developed in order to account for local item dependencies. The most common approach to LID in psychometrics is the *testlet approach* (Keller, Swaminathan, & Sireci, 2003; Thissen, Steinberg, & Mooney, 1989; Wilson, 1988; Wilson & Adams, 1995) as will be discussed in the following. Examples are sets of items associated with the same stimulus or item stem, such as the same reading paragraph, and sets of items relying on the same type of knowledge. Bao and Mislevy (2003) argue that such items are desirable, for example, in testing science performance, because they reflect real life situations. The referred item groups are denoted as *testlets* (Wainer, & Kiely, 1987), *item bundles* (Rosenbaum, 1988) or *superitems* (e.g., Ferrara, Huynh, & Baghi, 1997). The sum of the items scores on a testlet can be treated as a score on a polytomous item. Consequently, polytomous item response models can be applied such as the partial credit model (Masters, 1982), generalized partial credit models (Muraki, 1992), and graded response models (Samejima, 1969).

Apart from a regrouping of the items, such as in the testlet approach, three basic types of modeling approaches are described in the statistical literature (Tuerlinckx & De Boeck, 2004; Molenberghs & Verbeke, 2005): (1) random effect models (e.g., Bradlow, Wainer & Wang, 1999), (2) conditional models (e.g., Kelderman & Rijkes, 1994), and (3) marginal models (e.g., Ip, 2002), each of which will be described in the following.

1. Random effect models Person based random effects corresponding to item clusters can account for associations of items, and hence, they can be sources of item dependencies. Random effects can be introduced for subsets of items, when there is a natural clustering of items, such as item groups concerning the same reading paragraph or the same type of cognitive processes. In case of K such item clusters, K+1 design factors (one per cluster, and

a general one; or only K if all items belong to a cluster) can be introduced each with a corresponding random weight. This random effect approach was used by Bradlow et al. (1999) as an alternative approach to testlets, and it is described also by Scott and Ip (2002).

2. Conditional models The second type of item dependency models are conditional models which express the probability of a response on an item conditionally on the response on one or more other items. A prominent example of conditional models is the group of loglinear models with the items as factors in the loglinear design (Ip, Wang, De Boeck & Meulders, 2004; Kelderman & Rijkes, 1994). Tuerlinckx and De Boeck (2004) differentiate between recursive and non-recursive dependencies in conditional models, referring to structural equation modeling (Bollen, 1989). In case of recursive dependency, and given a particular ordering of the items, responses on items can affect the responses on items ordered after them, but not vice versa. For instance, learning effects during test taking produces a recursive type of item dependency with time as the natural basis of the ordering.

The most flexible recursive model allows for effects of all preceding responses on all following responses. Several simplifications of this model exist (Verhelst & Glas, 1993). The most common is a restriction to main effects of earlier responses (Tuerlinckx & De Boeck, 2004). Although in recursive models, the dependence of later item responses on earlier ones is a component of the model, and not vice versa, it can be shown mathematically that the response probabilities of earlier items are also affected by the responses on later items (Tuerlinckx & De Boeck, 2004; Verhelst & Glas, 1995).

In case of nonrecursive dependency, feedback loops are allowed among the items, and hence, mutual effects are possible. The fully parameterized model includes the effects of all item responses on all other item responses (Kelderman & Rijkes, 1994). A common restriction is to ignore interactions of a higher order than two-way interactions (Tuerlinckx & De Boeck, 2004).

The conditional model approach and the random-effect approach suffer from some common problems (Tuerlinckx & De Boeck, 2004). The marginal probability of item responses depends on the dependency parameters, and the item difficulty parameter loses its natural interpretation.

3. Marginal models The marginal model approach is a way to deal with these problems. Marginal models do have reproducible marginals, which means that the marginal probability of a correct response does not depend on LID. Examples are the multivariate probit model (Ashford & Sowden, 1970; Lesaffre & Molenbergs, 1991) and Ip's (2000, 2001, 2002) hybrid models for item dependencies.

Generalized estimating equations (GEE; Liang & Zeger, 1986) form a specific approach to marginal modeling. GEE models were developed for extending the Generalized Linear Models to correlated data where estimating equations are generalizations of other estimating equations (Hardin & Hilbe, 2003) and concentrate on only certain aspects of the data such as means and covariances. Alternating Logistic Regression (ALR; Carey, Zeger, & Diggle, 1993; Liang, Zeger, & Qaqish, 1992) is a specific GEE variant which combines a marginal logistic regression for the mean structure with a logistic regression for the association structure.

Item clusters can easily be introduced in this structure. Ideally, the marginal probabilities are independent of the association structure and the item difficulty parameters have their natural interpretation, but depending on the approach, the means and associations put restrictions on one another. The disadvantage of the marginal modeling approach is that it may become very computationally intensive if there are many items or if the model cannot be reformulated as a random effects model (Tuerlinckx & De Boeck, 2004).

The focus of this study is *LID of the random-effects type* which is created by a small additional dimension, independent of the substantive origin of the dependency (based on a

common item stem or overlap of cognitive processes, etc.). This is the type studied also by Bradlow et al. (1999). The Rasch model is used as the basic model. In the next section, several candidates for item dependency diagnosis are described which will be further used in a simulation study where their performance is assessed and compared. The approach of the study relies on the Rasch model, which is the model that corresponds best with using sum scores, as is commonly the case for tests.

The investigated dependency diagnostics

Three main types of diagnostics are distinguished in the following: pairwise diagnostics, clustering techniques and visual methods. Four *pairwise diagnostics* were selected: the modified Mantel-Haenszel test (Verguts & De Boeck, 2001), the likelihood ratio G^2 (Bishop, Fienberg & Holland, 1975), the standardized log-odds ratio difference (τ; Haberman, 1978), and Yen's Q_3 (1984). All four provide a significance test for the decision about dependency of single item pairs. The latter three were investigated in an earlier study by Chen and Thissen (1997), where it was found that Yen's Q_3 performs better for multidimensionality based LID (ULD) in data with an underlying 2PL and 3PL structure, and that Yen's Q_3 performs equally well as the other LID diagnostics when the LID is not dimensionality based (SLD). Chen and Thissen (1997) noted that the distribution of Yen's Q_3 is not as close to its theoretical distribution than of the likelihood ratio G^2.

In this study, beyond using these pairwise statistics for testing the LID for single item pairs, these pairwise statistics will also be used for deciding on the relative size of the pairwise dependencies. Furthermore, the pairwise statistics are also used as an input for the clustering methods and the grey-scale based visual approach.

As *clustering techniques*, DETECT (Kim, 1994; Stout, Habing, Douglas, Kim, Roussos & Zhang, 1996) and ADCLUS (Lee, 1999, 2001) were included. DETECT provides non-overlapping subgroups of items, while ADCLUS offers overlapping item clusters using similarities of the item pairs as input. The four above mentioned pairwise statistics will be used as input to ADCLUS. DETECT is applied on the raw data, hence one has not the choice to use these statistics. DETECT and ADCLUS not only extract the clusters, but also provide a decision on the number of clusters, so that the performance of these methods can be assessed on both, the recovery of the clusters, given the true number of clusters, and the recovery of the true number of clusters.

Finally, also a visual method will be investigated, with *grey-scale matrices* based on the four pairwise diagnostics. The inspection of the grey-scale matrices provides a rather quick and easy technique to explore and to visualize the underlying structure of data. In the following section, each of the above listed methods will be described in a more elaborated way.

Pairwise diagnostics

The modified Mantel-Haenszel test (MH) was proposed by Verguts and De Boeck (2001) as a distribution-free test for LID given the Rasch model as a null hypothesis. When it is suspected that two items (*i* and *i'*) are dependent, conditional on the latent trait, a contingency table can be created for each rest score group (the total score minus the score on item *i*). In each contingency table, the possible scores on item *i* define the rows and the scores on item *i'* define the columns. The observed frequencies of the item score patterns of the item pair in each rest score group are used to calculate the MH statistic, which is defined by Equation 1.

$$MH(i,i') = \frac{\left(\sum_{rs} O_{11rs} - \sum_{rs} E_{11rs}\right)^2}{\sum_{rs} \sigma^2(O_{11rs})}, \quad (1)$$

where O_{11rs} is the frequency in the rest score group *rs* of responding correctly to both items of the item pair (*i* and *i'*). E_{11rs} denotes the corresponding expected frequency, and $\sigma^2(O_{11rs})$ denotes the variance of the observed frequency in the rest score group *rs*. In general, O_{ab} (E_{ab}) is the observed (expected) frequency of item *i* having value *a* and item *i'* having value *b* at the same time. Furthermore,

$$E_{11rs} = (O_{11rs} + O_{10rs})(O_{11rs} + O_{01rs})/n_{rs},$$

$$\sigma^2(O_{11rs}) = E_{11rs}[(O_{10rs} + O_{00rs})/(n_{rs} - 1)][(O_{01rs} + O_{00rs})/n_{rs}],$$

and

$$n_{rs} = O_{11rs} + O_{10rs} + O_{01rs} + O_{00rs}.$$

Given LSI of item *i* and item *i'*, the MH is asymptotically χ^2 distributed with one degree of freedom. High MH values indicate item dependency.

The likelihood ratio G^2 (Bishop, Fienberg & Holland, 1975; Chen & Thissen, 1997) for item dependency is calculated here based on the conditional frequencies and their expectations as defined for the modified MH test:

$$G^2_{ii'} = -2\left(O_{11}\ln\left(\frac{E_{11}}{O_{11}}\right) + O_{10}\ln\left(\frac{E_{10}}{O_{10}}\right) + O_{01}\ln\left(\frac{E_{01}}{O_{01}}\right) + O_{00}\ln\left(\frac{E_{00}}{O_{00}}\right)\right), \quad (2)$$

where O_{ab} is the observed and E_{ab} is the expected frequency that item *i* has a value of *a* and item *i'* has a value of *b* at the same time, so that $O_{ab}=\Sigma_{rs}O_{abrs}$, and $E_{ab}=\Sigma_{rs}E_{abrs}$. In the way the test is defined here, it is also a distribution free test for LID given the Rasch model as a null hypothesis, just as the modified MH test. Under the model with LSI, the G^2 statistic is expected to be χ^2 distributed with one degree of freedom.

The *standardized log-odds ratio difference* (τ) is proposed by Haberman (1978). The formulas are given in Equations 3 and 4.

$$\tau_{ii'} = \frac{\tau_{ii'}^{obs} - \tau_{ii'}^{\exp}}{\sqrt{\sum_{i}\sum_{i'} 1/O_{ii'}}}, \tag{3}$$

where $\tau_{ii'}^{obs} = \ln\left(\frac{O_{11}O_{00}}{O_{10}O_{01}}\right)$, (4)

and $\tau_{ii'}^{\exp}$ is the expected log-odds ratio. O_{ab} and E_{ab} are defined as for the previous test. For independent item pairs, the τ statistic is expected to be normally distributed with zero mean and a variance of one (Chen & Thissen, 1997; Tate, 2003).

Yen's Q_3 (1984) is developed for investigating local item dependency based on an earlier statistic described by Kingston and Doran (1982). Q_3 is basically the correlation of the residuals for item *i* and for item *i'*. The residuals are calculated as the discrepancy between the observed values and corresponding expectations which in this study are estimated with the Rasch model. Q_3 is calculated as described by Equation 5 and Equation 6.

$$Q_{3ii'} = r_{d_{pi}d_{pi'}}, \tag{5}$$

where $r_{d_{pi}d_{pi'}}$ is the correlation of the residuals d_{pi} and $d_{pi'}$ for item *i* and item *i'* and

$$d_{pi} = y_{pi} - E(P(Y_{pi}|\theta_p)). \tag{6}$$

In Equation 6, y_{pi} is the response of person *p* on item *i* and $E(P(Y_{pi}|\theta_p))$ denotes the expected probability of person *p* to solve item *i* correctly, given the common dimension, θ_p. For the independence case, the Fisher's z transform of Q_3 is normally distributed with a zero mean and a variance of 1/(N-3) (Yen, 1984). In order to test the significance, the standardized Fisher's z transformation is used, dividing Fisher's z by the square root of 1/(N-3), so that a standard z-score is obtained. Q_3 has been used successfully in several studies (Chen &

Thissen, 1997; Fennessy, 1995; Yen 1993), suggesting it is a rather robust method. For the present study, the θ_p estimates were obtained by the PROC NLMIXED procedure of SAS for the random effects Rasch model, with quasi-Newton optimization and 20 quadrature points (SAS/STAT User's guide). For a decision on LID, the standardized Fisher's z transform of the Q_3 values were used.

Ferrara, Huynh, and Baghi (1997) noted that Yen's Q_3 method is somewhat limited by the fact that it requires a point estimate of the common ability parameter, and hence a relatively well fitting IRT model. The other three methods are nonparametric, although in principle, they can be derived also on a parametric basis. It is also worth noting that asymptotically, both, the Mantel-Haenszel test and the G^2 statistic are χ^2 distributed, while the standardized log-odds ratio difference (τ) and the Fisher's z transform of Q_3 are normally distributed. Besides, the first two LID statistics do not take the direction of the association into account, whereas the latter two do.

Clustering techniques

DETECT (Dimensionality evaluation to enumerate contributing traits; Kim, 1994; Stout, Habing, Douglas, Kim, Roussos & Zhang, 1996) is a method for revealing homogeneous item subgroups that represent a separate dimension. Therefore, it can in principle also be used for detecting the type of LID focused on this study. It is a method based on the conditional covariance of the item pairs (conditional on the general dimension).

A simple structure, or in other words, non-overlapping item clusters, is a requirement for DETECT (Stout, Habing, Douglas, Kim, Roussos & Zhang, 1996). A simple structure is realized when non-overlapping item subgroups can be identified *and* the items within subgroups measure the same ability (Tate, 2003; Zhang & Stout, 1999). The method is

expected to give a good estimate of the item clusters also for an approximate simple structure. DETECT provides an index for simple structure, called the R index. When its value is higher than 0.8, then an approximate simple structure can be assumed, and consequently, the DETECT index should give an accurate indication of the dimensionality of the data, and the accompanying item partition is assumed to be an accurate indication of the item clusters.

The DETECT software also offers cross-validation (Zhang & Stout, 1999). In the cross-validation procedure, one half of the data is used to obtain a partition of the items, which is then further used for the second subset of the data, providing the cross-validated DETECT index. In general, cross-validation must be recommended unless the sample size is too small to divide the data set into two parts.

Additive clustering (ADCLUS) can be used even if the simple structure is not realized, because for ADCLUS the item clusters overlap. Here, ADCLUS will be used as developed by Lee (1999, 2001). The version of the program used here, produces an item cluster matrix with the requested number of possibly overlapping item clusters. ADCLUS requires similarity measures as input, and the four types of pairwise statistics will be used for that purpose. For a decision on the number of clusters, a scree test will be applied on the weights of the item clusters. From a former study (Balazs, Schepers, & De Boeck, 2006; see Chapter 3) it could be concluded that ADCLUS may be used effectively to extract item clusters, even when theoretically the clustering results may depend on the marginals (see a remark in Chapter 3).

Visual methods

Visual inspection of grey-scale matrices (Tuerlinckx & De Boeck, 2004) is another way to detect dependencies. The result of the association measures can be represented with grey-scale matrices, darker cells indicate higher values. The items form the rows and also the

columns of the grey-scale matrix, so that both the upper and lower triangle shows the pairwise LID. When the observed grey-scale matrix resembles quite well to the matrices derived from the model, it is concluded that there is no LID in the data, but when clear differences are found, they indicate LID.

If the true model of the data structure is the Rasch model, one would expect uniform grey-scales matrices. A clear differentiation within an observed grey-scale matrix indicates LID and where it occurs. The method is used for the modified MH statistic, the likelihood ratio G^2, the standardized log-odds ratio, and Yen's Q_3.

Simulation study

A simulation study was set up in order to compare the above-mentioned methods for detecting dimensionality based LID, while LID is defined in comparison to the Rasch model. The data structure was a random weights LLTM (Rijmen & De Boeck, 2002) for 24 items and sample sizes of 100, 500, and 1000. For all data sets, a general underlying dimension was used, and for half of the data sets a cluster of three items defines a second dimension, while for the other half of the data sets an additional second cluster of three items defines a third dimension. See Equation 7 for the first half of the data sets and Equation 8 for the second half of the data sets: The model of Equation 7 leads to a two-dimensional data structure, while the model of Equation 8 leads to a three-dimensional data structure. In the former, there is one LID cluster of three items, and in the latter there are two such clusters:

$$\text{logit}\left(P_{pi}(Y_{pi}=1|\theta_p,\beta_{p1})\right)=\theta_p+\beta_{i0}+\beta_{p1}X_{i1}, \quad (7)$$

$$\text{logit}\left(P_{pi}(Y_{pi}=1|\theta_p,\beta_{p1},\beta_{p2})\right)=\theta_p+\beta_{i0}+\beta_{p1}X_{i1}+\beta_{p2}X_{i2}, \quad (8)$$

where $\left(P_{pi}(Y_{pi}=1 \middle| \theta_p, \beta_{p1}, \beta_{p2}) \right)$ is the success probability of person p for solving item i correctly, modeled as a function of two item covariates (X_{i1} and X_{i2}). θ_{p} is the common latent dimension. X_{i1} and X_{i2} are the covariates for item i (i=1,..., I), and β_{p1} and β_{p2} are the associated random weights for person p. The formulation in Equation 7 is analogous for the case of one cluster of three items. The distributions of $(\theta_{\mathrm{p}}, \beta_{p1})$, and of $(\theta_{\mathrm{p}}, \beta_{p1}, \beta_{p2})$, are multivariate normal with mean zero, and the variance of θ_{p} is one.

For all data sets, X_{i1}=1 for items 4, 5, and 6, and for half of the data sets, X_{i2}=1 for items 19, 20, and 21, while X_{i1}=0 and X_{i2}=0 in all other cases. The variances of the covariate based dimension(s), β_{p1} and β_{p2}, are either 0.5, or 1, or 2. The correlation of all three random effects is either 0 or .5. The design is fully crossed in the simulation study. Three sample sizes (100, 500, 1000), two dimensionality values (two, three), three variance values (0.5, 1, 2) and two correlation values (0, .5) yield 36 design cells. For each cell in the design, 10 data sets were generated.

A larger variance of the minor dimension(s) increases the LID in the corresponding item pairs, and the correlation of the minor dimension(s) with the general dimension decreases the amount of LID in the corresponding item pairs. On the other hand, the sample size and the dimensionality do not affect the amount of LID in the item pairs. Therefore, an ideal LID diagnostic for dependent pairs would show sensitive to the variance and the correlation of the dimensions, but not to the sample size, except as far as the statistic also reflects the power of the method.

Results

Before discussing the results, it is important to note that the data structures of the simulation study lead to three different types of item pairs. *Type 1* concerns pairs within the same LID cluster, called the *dependent item pairs*. This is the kind of pairwise dependence focused on in this study. *Type 2* concerns pairs with items belonging to different clusters. This type is of importance only if the two item clusters are correlated, and it is by definition restricted to data sets with two LID clusters. *Type 3* concerns pairs with one or two items not belonging to any LID cluster, called the *independent item pairs*.

Pairwise statistics

First, the performance of the pairwise statistics is investigated in terms of hits and false alarms, with $\alpha = .05$ to determine the statistical significance. The hit rates and false alarm rates are estimated on the basis of the Type 1 and Type 3 pairs, respectively. Second, the effects of the design factors have on the values of the pairwise diagnostics are summarized (in terms of the mean values of the statistics), and the relative performance of the diagnostics is evaluated (in terms of percentiles of the dependent item pairs indicating their place in the distribution of all item pairs). Finally, the methods are compared with one another.

The modified Mantel-Haenszel test

The Mantel-Haenszel test could not be calculated for 12 item pairs in total, all occurring in data sets with sample size 100. The reason for the problem was that the $\sigma^2(O_{11rs})$ term became zero. For these 12 cases out of 99360 cases, no data are available.

Applying a significance level of .05 on the Mantel-Haenszel values, 54.5% hits and 5.6% false alarms are observed. The hit and false alarm rates (percent of significant dependent

and independent item pairs, respectively) vary depending on the design factors; see Table 1. With an increasing sample size, the hit rate improves considerably, up to 81.7% for N=1000, and the false alarm remains slightly above the theoretical 5%. As expected, the hit rate also increases with the variance, and it decreases with the correlation. The dimensionality does not seem to have much of an effect. The false alarm rate seems to increase slightly as a function of the variance, but not really as a function of the sample size.

Insert Table 1 about here.

When the means of the MH values for the dependent item pairs are looked at, it must be concluded that they vary as a function of the same design factors as the hit rates; see Table 2. For dependent item pairs, the values are higher for a larger sample size, for a larger variance, and for a smaller correlation, whereas the difference is only small between two and three dimensions. For the independent pairs, the values do not vary with the design factors, as expected.

Insert Table 2 about here.

The relative performance of the MH test is evaluated through the minimum percentile of MH values for dependent item pairs. For each MH test value, a corresponding percentile value defines the percent of the distribution with a value equal to or below to it. As shown in Tables 5 and 6, for two and three-dimensional data, the minimum percentiles are high if N is 500 or larger and the variance is 2. Assuming there are no ties among the dependent item

pairs with the smallest dependency value, ideally, the minimum percentile is 99.3 for two-dimensional data, and 98.2 for three-dimensional data. The values of 99.3 and 98.2 are actually also established when N=1000, the variance is 2 and the correlation is zero.

Insert Table 5 about here.

Insert Table 6 about here.

Likelihood ratio G^2

Likelihood ratio G^2 estimates have been obtained for all item pairs. The G^2 test resulted in 48.5% hits and 2.2% false alarms. Both the hit and the false alarm rates are lower than those obtained with the MH test. The hit and false alarm rates vary depending on the design factors; see Table 1. With an increasing sample size, the hit rate improves considerably, up to 77.0% for N=1000, and the false alarm rate increases slightly but even for N=1000, it does not approach the theoretical 5%. As expected, the hit rate, but not the false alarm rate, also increases with the variance, and it decreases with the correlation. The dimensionality does not seem to have an effect.

When the means of the likelihood ratio G^2 values for the dependent item pairs are looked at, it must be concluded that they vary as a function of the same design factors as the hit rates; see Table 2. The dependency values of the dependent pairs are smaller than those of the MH test, and they increase with the sample size and with the variance, and they decrease

with the correlation, whereas the difference is only small between two and three dimensions. The means for the independent pairs seem to increase slightly with a larger sample size, but do not vary with the rest of the design factors.

The relative performance of the likelihood ratio G^2 statistic seem to be about the same as the relative performance of the MH test both for two- and three-dimensional data, as it can be seen in Tables 5 and 6.

Standardized log-odds ratio difference (τ)

Standardized log-odds ratio difference (τ) estimates have been obtained for all item pairs. The results of the method are very similar to those obtained with the likelihood ratio G^2 method. The test resulted in 46.7% hits and 2% false alarms. As Table 3 shows, the percentages of hits and false alarms are smaller than for the MH test and slightly smaller than those for the likelihood ratio G^2 method. The hit rates increase with the design factors as expected, but the false alarm rates do as well, although only to a small extent.

For dependent pairs, also similar effects of the design factors on the statistic were found as for the MH test and the likelihood ratio G^2 test (Table 4). A remarkable finding regarding the independent pairs is that the value of the dependence statistic deviates slightly more from zero in the negative direction with increasing sample size and increasing variance.

As can be seen in Tables 5 and 6, the relative performance of τ follows the pattern of the MH test and likelihood ratio G^2, but, in general, the percentiles are higher indicating a better performance.

Insert Table 3 about here.

Insert Table 4 about here.

Yen's Q_3

The Q_3 estimates have been obtained for all item pairs. The statistical significance of the standardized Fisher's z transformation of the Q_3 values was used to decide about the LID of the item pairs. With the Q_3 test, 45.8% hits and 14.6 % false alarms were found; see Table 3. The results regarding the overall hit rate are very similar to those obtained with standardized log-odds ratio difference (τ), but the false alarm rates are the highest when compared among the four pairwise diagnostics.

It is interesting to note that although the Q_3 values are not a function of a sample size, the standardized Fisher's z transform of the Q_3 statistic for dependent pairs becomes a function of the sample size indicating that the power of the method is a function of the sample size. Similarly, the false alarms rate also seems to increase with the sample size, further away from the theoretical 5%.

The effects of the design factors on the Q_3 values are similar to the effects obtained for the standardized log-odds ratio difference (τ) method. The high negative values of the Q_3 for independent pairs can explain the higher false alarm rates.

The relative performance follows the pattern of the other three statistics, but the minimum percentiles are higher than for any of the other three.

A comparison of the pairwise statistics

In general, the number of dependent item pairs is underestimated with all four methods. As could be expected, the hit rate clearly improves with an increasing sample size, with an increasing variance of the random effects, and when the random effects are not correlated. At the same time, the dimensionality of the data did not remarkably affect the performance of the pairwise diagnostics, but in general, the hit rate results are slightly worse for three dimensions (two LID clusters).

The false alarm rate is lower than the theoretical 5% for G^2 and τ, unlike for the MH test where it is slightly larger and for Q_3, where it is much larger from a sample size of 500 on. For G^2 and τ, the false alarm rate increases as a function of the sample size, but does not reach 5% even for N=1000. From a comparison of the hit rates and false alarms, the MH test has the best performance of the four diagnostics. In general, a sample size of 500 does not seem sufficient for an overall satisfactory performance, while for a sample size of 1000, the results seem more reasonable.

The effect of the design factors on the values of the four statistics for the dependent items parallels the results for the hit rates. Higher values are found as a function of the sample size and variance, and lower values are found for zero correlations than for a .5 correlation. The test values, and hence the power of the test, increase with the sample size. Also for independent item pairs, the sample size seems to have a sizable effect, although much smaller than for dependent pairs.

The pairwise methods are also investigated as potential relative criteria, as an ordering criterion for the dependency. A sample size of 100 does not seem to be sufficient for any of the methods to obtain high percentiles for the smallest dependency value among the dependent pairs. When the sample size is reasonably large (at least 500) and the variance is also large (as large as 2), all methods work reasonably well, but perfection is obtained only with a sample size of 1000, a variance of 2 and a zero correlation. When the minimum

percentiles of the dependent item pairs are considered as a criterion for the relative performance, Yen's Q_3 and the standardized log-odds ratio difference (τ) seem to perform best.

As a criterion for decision making, the modified Mantel-Haenszel test had the best performance, while as a relative criterion Yen's Q_3 and the standardized log-odds ratio difference (τ) had the best performance, and Yen's Q_3 seems slightly superior. In general, a sample size as large as 1000 seems required in order to obtain reasonable results, even for the best performing statistics.

Clustering methods

The performance of the clustering methods is assessed in terms of the cluster content and in terms of the decision on the number of clusters.

DETECT

The DETECT procedure was applied to the simulated data, both, with and without cross-validation. In the DETECT procedure, the user has to specify the maximum number of item clusters to be extracted. The evaluation is made in two steps. First, in order to focus on the effectiveness of DETECT in revealing the content of the item clusters, the true number of clusters is considered as given. Therefore, a maximum of two item clusters is chosen for the two-dimensional data sets since two clusters can be differentiated in the two-dimensional space (one cluster of three items and one cluster with all the other items), and a maximum of three clusters is chosen for the three-dimensional data sets, since three clusters can be differentiated (two clusters of three items and one cluster with all other items). Hence, the maximum was determined as the true number (for an explanation of the issue, see further).

Second, it will be evaluated whether DETECT can actually find the true number of clusters. In order to do so, a higher maximum number of clusters was chosen (five clusters), to give the freedom of an overestimation of the true number of clusters.

Considering either the result of DETECT with cross-validation or the results without it, the R index value (the index of simple structure) never exceeded a value of 0.8. The mean of the R index values were 0.05 and 0.45, with and without cross-validation, respectively. Because a lack of approximate simple structure was indicated for all data sets, no further conclusions can be drawn regarding the data structure. This result is not surprising because the data were not generated from a simple structure, all items being linked to a common dimension, and additionally, either one set of three items or two sets of three items are linked to a second dimension.

A further problem for the evaluation of DETECT in terms of cluster extraction is that one can determine the maximum number of clusters but not a given number of clusters to be extracted. And even if the desired number of clusters is extracted, when the DETECT index value is smaller than a critical value (0.1 or 0.2), it is an indication that the number of clusters is actually smaller than the extracted one. Without cross-validation and with a critical value of 0.1, the true number of clusters is indicated for 100% and 95.6% of the two- and three-dimensional data, respectively. Without cross-validation and with a critical value of 0.2, the correct number of clusters was indicated for 74.4% and 93.4% of the two- and three-dimensional data sets, respectively. Furthermore, with cross-validation, DETECT often indicates a smaller number of clusters than the desired one. With cross-validation and with a critical value of 0.1, the correct number of clusters is indicated for 12.8% of the two-dimensional and 28.9% of the three-dimensional data sets. With cross-validation and with a critical value of 0.2, DETECT indicates the correct number of clusters for 3.3% of the two dimensional and 1.1% of the three-dimensional data sets. This implies that the first evaluation

step, focusing on cluster membership given the true number of clusters, is problematic, unless no cross-validation is used and a critical value of 0.1 is chosen, because then in almost all cases the maximum was also the true number. However, it is clear from the cases where the true number of clusters is selected, that the clusters have an about equal size, so that their content is certainly not in agreement with the true clusters.

Nevertheless, the results indicate that the DETECT values are a function of the same aspects of the minor dimension(s) as the pairwise statistics are. When the DETECT procedure without cross-validation is considered, for data with a variance of 0.5, 1 and 2, the mean DETECT values are 0.35, 0.36 and 0.41, respectively. The mean DETECT value is 0.39 for uncorrelated and 0.36 for correlated data. The mean DETECT value is 0.33 for the two-dimensional data sets and 0.42 for the three-dimensional data sets. The cross-validated DETECT index is a function of the same design factors. For data with a variance of 0.5, 1 and 2, the mean DETECT values are 0.03, 0.03 and 0.07, respectively. The mean of the DETECT values is 0.03 for correlated data, and it is 0.05 for uncorrelated data. The mean of the cross-validated DETECT value is .01 for two-dimensional and 0.07 for three dimensional data. This latter effect is different from the effect on the pairwise statistics.

For the second step of the evaluation, DETECT was allowed to choose the number of clusters. The maximum number of clusters was set to five. In principle, the DETECT values could not be interpreted in this step of the evaluation either, because of lack of approximate simple structure. In order to pursue the investigation of DETECT, the role of the R index was ignored for this second step. With a maximum of five clusters, DETECT may extract two to five clusters. However, if the DETECT value does not reach 0.1 as described in the manual (The William Stout Institute for Measurement, 2003) or 0.2 as it is recently suggested (van Abswoude, van der Ark, & Sijtsma, 2004), the conclusion must be that there is only one (common) dimension, and that there is not good evidence for another source of heterogeneity.

The two-dimensional data defines two clusters, while the three-dimensional data defines three item clusters. When the DETECT procedure was applied without cross-validation, the results were the same for the two possible critical values (0.1 or 0.2). In case of two-dimensional data, two, three, four and five clusters were indicated for 1.7%, 35.5%, 54.4% and 8.3% of the data sets, respectively. In case of three-dimensional data sets, two, three, four and five clusters were indicated for 1.7%, 33.8%, 51.1% and 13% of the data, respectively. Therefore one can conclude that the number of clusters is mostly overestimated by DETECT without cross-validation.

When the cross-validated DETECT is applied to two-dimensional data, with a critical value of 0.1, 87.2% of the data sets are indicated as unidimensional, and the remaining 12.8% of the data sets are indicated having two item clusters. Applying a critical value of 0.2, 96.7% of the data sets are indicated to be unidimensional, and 3.3% of the data sets are indicated to have two item clusters.

When the DETECT procedure is applied with cross-validation and with a critical value of 0.1 to three-dimensional data, 57.8% of the data sets are shown as unidimensional, 5.6% are indicated as having two item clusters and 36.7% as having three item clusters. When the same procedure is used with a critical value of 0.2, 86.7% of the three-dimensional data sets are shown as unidimensional, 2.8% were indicated as having two item clusters and 10.5% as having three item clusters.

Consequently, the DETECT procedure with cross-validation seem to underestimate the number of item clusters, which is not surprising given the small size of the additional dimensions. The results are slightly better with a critical DETECT value of 0.1.

In sum, when following the rules, the DETECT value should not be interpreted, because the R value indicates an absence of approximate simple structure. When the R value is ignored, and the true number of item clusters is set as the maximum number of clusters, the

DETECT value obtained without cross-validation is quite successful in detecting the multidimensionality of the data, but the cross-validated DETECT is not, although theoretically the latter should have a better performance. However, DETECT does not seem able to recover the content of the clusters. Concerning the recovery of the number of clusters, the DETECT procedure without cross-validation seems to overestimate the number of clusters, and with cross-validation it seems to underestimate the number of clusters.

ADCLUS

The ADCLUS procedure was applied on all four pairwise statistics. In the following, the performance of ADCLUS is investigated both in terms of recovering the cluster content and in terms of detecting the true number of item clusters. In order to investigate the former, the true number of clusters is extracted (one or two beyond the general cluster), and in order to check whether the true number can be found, five clusters were extracted, so that a scree test can be used to select a number of clusters.

The adjusted Rand index (Hubert & Arabie, 1985; see also, Fowlkes & Mallows, 1983; Rand, 1971; Yeung, & Ruzzo, 2006), a measure of agreement of two partitions, was calculated for comparison with the true structure. The higher the adjusted Rand index value is, the more similar the clusters are; the adjusted Rand index has an upper bound of 1. The adjusted Rand index is calculated in all possible combinations of the extracted and true clusters, and the highest adjusted Rand index value is chosen among the obtained Rand index values.

First, the performance of the methods in terms of the recovery of the cluster content is shown in Table 7. Based on the overall mean of the adjusted Rand index values, the likelihood G^2 and standardized log-odds ratio difference (τ) used for ADCLUS were the best

for item cluster extraction, the adjusted Rand indices being .38 and .36, respectively. The performance varies as a function of the design factors in the same way as before. For data with a sample size of 1000, with a variance of 2, and with a correlation of zero, the mean of the adjusted Rand index values was 0.96 for the standardized log-odds ratio difference (τ).

Insert Table 7 about here.

Second, for a decision on the number of item clusters, a scree test was applied on the weights of the ADCLUS derived five item clusters. Following a principle explained by Ceulemans and Van Mechelen (2005), a discrepancy difference measure is calculated per weight: the discrepancy with the following weight (current weight-following weight) is subtracted from the discrepancy with the previous weight (previous weight-current weight) while the weights are in decreasing order. The first measure is therefore determined for the second weight and the last one is determined for the fourth weight (given a maximum of five clusters). The discrepancy difference measure can be mapped against the cluster number, and based on the elbow criterion a number of clusters can be selected. The weight with the maximal discrepancy difference is selected (as the elbow), and the number of item clusters is determined as the order position of the selected weight minus one. In this way, the possible outcomes for five clusters are one to four item clusters.

The percentages of correct decisions on the number of clusters as a function of the design factors, are shown in Table 8. For the MH and G^2 statistics, the effects do not follow the expectations, but for τ and Q_3 statistics they do. It can be seen that ADCLUS applied on the standardised log-odds ratio difference (τ) performed somewhat better than ADCLUS applied on the other measures. For data with a sample size of 1000, with a variance of 2 and

with a correlation of zero, the percentage of correct decisions for the standardized log-odds ratio difference (τ) was as high as 100% (the combination is not shown in Table 8).

Insert Table 8 about here.

In a paper by Balazs, Schepers, and De Boeck (2006 ; see Chapter 3), ADCLUS is used with the simple unconditional log-odds ratio. One may wonder how it performs in comparison with the four statistics that are used here. As one can note in Tables 7 and 8, for the recovery of the cluster content, the log-odds ratio seems to be by far the better method, but for a decision on the number of clusters, the conditional approach with the standardized log-odds ratio difference (τ) yields a somewhat higher number of correct decisions. The good performance of the log-odds ratio based methods is not surprising given the additive basis of ADCLUS and of the log-odds ratio (see Chapter 3).

In sum, the original method with the log-odds ratios and the method with the standardized log-odds ratio difference (τ) as input to ADCLUS performed the best. One should note that the log-odds ratios were not calculated conditionally on the general dimension, while the four statistics were indeed. Therefore, the performance of ADCLUS applied on the log-odds ratios may be improved for determining the number of true clusters by conditioning on the sum scores as for the standardized log-odds ratio difference (τ).

Visual methods

As an illustration that grey-scale matrices may provide a proper way to inspect and to visualise the dependency structure of data, four design cells were selected to construct grey-scale matrices. For all four pairwise statistics, the same four randomly selected data sets were

used to illustrate the results in each of the four cells. The values of the pairwise statistics were standardized in order to obtain comparable grey-scale values for each data set. The four selected design cells are all cells with three-dimensional data and with 0.5 as a correlation of the dimensions (1) N=500, variance is 0.5, (2) N=500, variance is 2, (3) N=1000, variance is 0.5, (4) N=1000, variance is 2. In this way, there is for both sample sizes (N=500 and N=1000) a relatively easy condition (with a variance of 2), and a relatively difficult condition (with a variance of 0.5). These four cells are indicated in bold in Table 6. Figure 1 shows the expected grey-scale matrices given the true structure, whereas Figures 2 to 5 show the results for the four cells, based on four representative data sets for each.

In this study, only the upper triangle of the panels is used to represent the item pairs. Therefore, the most left column of grey squares refers to item 1, while the most right column refers to item 23. The most upper row of grey squares represents the values for item 24, while the most bottom row represents the values for item 2. The darker squares represent higher values.

Insert Figure 1 about here.

As can be seen in Figure 1, two triplets of items should stick out with a higher dependence (Type 1 pairs), because three-dimensional data sets are considered. In case of a correlation between the two minor dimensions, also nine other of the item pairs should be somewhat darker (Type 2 pairs). Finally, the remaining item pairs (Type 3) should show uniform grey-scale values.

Insert Figure 2 about here.

Insert Figure 3 about here.

In the Figures from 2 to 5, the main columns refer to the four types of pairwise statistics as a basis of the matrices. The matrices are labelled MH, G2, TAU, and Q3 denoting the modified Mantel-Haenszel test, likelihood ratio G^2, the standardized log-odds ratio difference (τ) and Yen's Q_3, respectively. The rows correspond to four different data sets (representative ones), numbered 1 to 4.

Insert Figure 4 about here.

Insert Figure 5 about here.

Comparing Figures 2 and 3 to Figures 4 and 5, respectively, it seems that the dependent item pairs can be distinguished slightly better based on the grey-scale matrices for design cells 3 and 4, which contain data with a larger sample size (1000), than for design cells 1 and 2, which contained data with a sample size of 500. The performance is worse for a sample size of 100 than for a sample size of 500, but not shown in any of the figures.

When Figure 2 is compared to Figure 3, and Figure 4 is compared to Figure 5, it is clear that the larger the variance of the minor dimension is, the easier it is to distinguish the dependent item pairs from the independent item pairs based on the grey-scale matrices. This effect is clearly larger than the effect of the sample size. None of the four statistics provides a very good basis for a visual inspection if the variance of the minor dimension(s) is 0.5.

The cell that differentiates best between dependent and independent pairs is the one with N=1000, and with a variance of 2, as may be derived from Table 6.

Although it has not been demonstrated here, but in agreement with the results obtained for the pairwise statistics as relative criteria, the results are slightly better for two-dimensional data than for three-dimensional data, and slightly better for data with uncorrelated dimensions than for data with correlated dimensions.

In sum, grey-scale matrices created from the values of the pairwise item dependency diagnostics can be effectively used to inspect and visualize the data structure when the sample size is sufficiently large and the variance of the minor dimension(s) is also sufficiently large. No conclusions are drawn on the four statistics, because only four data sets in each design cell are discussed. The grey-scale matrix approach is discussed here only for illustrative purposes.

Conclusions

In the present study, four pairwise diagnostics were compared on their performance as absolute and relative criteria to detect LID, and they were also used as inputs to a cluster algorithm for overlapping structures to find sources of LID and as inputs to a visual technique. Additionally, another clustering algorithm, DETECT was also tried out.
From the investigated criteria, the modified Mantel-Haenszel test performed the best as an absolute criterion. As relative criteria, Yen's Q_3, and the standardized log-odds ratio

difference (τ) performed the best. In general, a large sample size, and independent and rather strong dimensions underlying the dependent pairs are required in order for these dimensions to be detected by the statistics.

As a clustering methods, DETECT cannot be a successful approach for the type of investigated data with minor additional dimensions, because of the lack of approximate simple structure. On average, ADCLUS did not perform very well either. However, the performance clearly varied as a function of the design factors. The performance was best when ADCLUS was applied to the standardized log-odds ratio difference (τ) and to the simple log-odds ratio.

Finally, also a visual method can be used for LID detection with the investigated pairwise diagnostics. Apparently, also for this method a large sample size and large variance are required.

References

Ackerman, T. (1987). *The robustness of LOGIST and BILOG estimation programs to violations of local independence.* ACT research report series, 87-14. Iowa City, IA: American Collefe Testing.

Anastasi, A (1961). *Psychological testing.* New York: Macmillan.

Ashford, J. R., & Swoden, R. R. (1970). Multivariate probit analysis. *Biometrics, 26,* 535-546.

Balazs, K., Schepers, J., & De Boeck, P. (unpublished manuscript). The detection of hidden item properties and dimensionality.

Bao, H., & Mislevy, R. J. (2003). An application of the multidimensional random coefficients multinomial logit item bundle model.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambrige, MA: MIT Press.

Bollen, K. A. (1989) *Structural equations with latent variables*. New York: Wiley.

Bradlow, E. T., Wainer, H., & Wang, X. (1999) A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Carey, V. J., Zeger, S. L., & Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika, 80*, 517-526.

Ceulemans, E., & Van Mechelen, I. (2005). Hierarchical classes models for three-way three mode binary data: Interrelations and model selection. *Psychometrika, 70*, 461-480.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Embretson, S. E., & Reise, S. P. (2000*). Item response theory for psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Fennessy, L. M (1995). *The impact of local dependencies on various IRT outcomes.* Unpublished doctoral dissertation. Univercity of Massachusetts at Armherst. [Dissertation Abstracts International]

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale assessment. *Applied Measurement in Education, 10,* 123-144.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGaw-Hill.

Haberman, S. J. (1978). Analysis of qualitative data: Vol. 1. Introductory topics. New York: Academic Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, NJ: Sage.

Hardin, J. W. & Hilbe, J. M. (2003). Generalized estimating equations, Chapman & Hall

Hoskens, M., & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement. 32*, 364-384.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local item dependencies among test items. *Psychological Methods, 2*, 261-277.

Hoskens, M., & De Boeck, P. (2001). Multidimensional componential IRT model. *Applied Psychological Measurement, 25,* 19-27.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification, 2-3,* 193-218.

Ip, E. H. (2000). Adjusting for information inflation due to local dependence in moderately large item clusters. *Psychometrika, 65,* 73-91.

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

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revised. *Psychometrika, 67,* 367-385.

Ip, E. H., Wang, Y. J., De Boeck, P., & Meulders, M. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility *Psychometrika, 69*, 191-216.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357 373.

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

Keller, L. A. , Swaminathan, H., & Sierci, S. G. (2003). Evaluating scoring procedures for context-dependent items sets. *Applied Measurement in Education, 16,* 207-222.

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crum's criticism. *The Journal of Educational Psychology, 15,* 193-204.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. (Doctoral dissertation, University Illinois at Urbana Campaign). *Dissertation Abstracts International, 55-12B*, 5598

Kingston, N. M., & Doran, N. J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (ETS Research Report 82-12). Princeton, NJ: Educational Testing Service. In Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using itemresponse theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Lee, M. D. (1999). An extraction and regularization approach to additive clustering. *Journal of Classification, 16,* 255-281.

Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology, 45,* 131-148.

Lesaffre, E. & Molenberghs, G. (1991). Multivariate probit analysis using generalized linear models. *Biometrika,* 73, 13-22.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalizes linear models. B*iometrica, 73*, 13-22.

Liang, K.-Y. & Zeger, S. L. & Oaquis, B. (1992). Multivariate regression models for categorical data. *Journal of the Royal Statistical Society. Series B. 54*, 3-40.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

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

Molenberghs, G. & Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer-Verslag.

Muraki, E. (1992) Ageneralized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association, 66,* 846-850.

Rijmen. F. & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement. 26,* 271-285.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349-359.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*, 1-100.

SAS/STAT User's guide [Software manual]

Scott, S. & Ip, E. (2002). Empirical Bayes and item clustering effects in a latent variables hierarchical model: A case study from the National Assessment of Educational Progress. *Journal of the American Statistical Association, 97,* 409-419.

Sireci, S., G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet based tests. *Journal of Educational Measurement, 28,* 237-247.

Spray, J., & Ackerman, T. (1987). *The effect of item response dependency on trait or ability dimensionality.* ACT Research Report Series, 87-10. Iowa City, IA: American College Testing.

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

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological measurement*. 27, 159-203.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace line for testlets: A use of multiple categorical response models. *Journal of Educational Measurement, 26,* 247-260.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp.560-620). Washington, D. C. American Council on Education.

Tuerlinckx, F., & De Boeck, P. (1998). Modeling local item dependencies in item response theory. Psychologica Belgica, 38, 61-82.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.) *Explanatory item response models*. (pp.289-316). Springer-Verlag New York.

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

Verguts, & De Boeck (2001) Some Mantel-Haenszel test of Rasch model assumptions. *British Journal of Mathematical and Statistical Psychology, 54,* 21-37.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*, 395-415.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Moleanaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 215-238). New York: Springer-Verslag.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computer adaptive testing: A case for testlets. *Journal of Educational measurement, 24,* 185-201.

The William Stout Institute for Measurement. (2003). DETECT manual.

Wilson, M. (1988). Detecting and implementing local item dependence using a family of Rasch models. *Applied Psychological Measurement, 12*, 353-364.

Wilson, M., & Adams, R. A. (1995).Rasch models for item bundles. *Psychometrika, 60,* 181 198.

Yen, W. M. (1984). Effects of local item dependence on the fit and equation performance of the three-parameter logistic model. *Applied Psychological measurement, 2,* 125-145.

Yen, W. M. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Yeung, K. Y. & Ruzzo, W. L. (05. 01. 2006) Details of the adjusted Rand index and clustering algorithms supplement to the paper " An empirical study on principal component analysis for clustering gene expression data", downloaded from: http://faculty.washington.edu/kayee/pca/supp.pdf

Zensky, A. L., Hambleton, R. K., & Sireci, S., G. (submitted). Effect of local item dependence on the validity of IRT item, test, and ability statistics. *Effects of Local Item Dependence.*

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

Table 1

The percent of significant dependent and independent item pairs as a function of the design factors (α=0.05)

		MH		G^2	
Design factor	Value	Dependent	Independent	Dependent	Independent
Sample size	100	22.6	5.6	14.8	1.6
	500	61.1	5.3	53.7	2.3
	1000	81.7	5.8	77.0	2.7
Variance	0.5	28.9	5.0	22.2	2.1
	1	57.4	5.3	49.6	2.1
	2	79.1	6.4	73.7	2.3
Correlation	0	62.1	5.4	57.2	2.4
	0.5	48.2	5.7	39.9	2.0
Dimensionality	2	57.8	5.6	50.4	2.1
	3	53.8	5.6	47.6	2.2
Overall		54.5	5.6	48.5	2.2

Table 2

The mean of the pairwise statistics for dependent and independent item pairs as a function of the design factors

		MH		G^2	
Design factor	Value	Dependent	Independent	Dependent	Independent
Sample size	100	2.5	1.0	1.7	0.6
	500	9.0	1.0	6.7	0.7
	1000	18.0	1.1	13.7	0.8
Variance	0.5	3.2	1.0	2.6	0.7
	1	7.4	1.0	5.5	0.7
	2	18.8	1.1	14.1	0.7
Correlation	0	12.1	1.0	9.7	0.7
	0.5	7.6	1.0	5.1	0.7
Dimensionality	2	10.3	1.0	7.8	0.7
	3	9.6	1.0	7.1	0.7
Overall		9.8	1.0	7.4	0.7

Table 3

The percent of significant dependent and independent item pairs as a function of the design factors (α=0.05)

		τ		Q_3	
Design factor	Value	Dependent	Independent	Dependent	Independent
Sample size	100	11.9	1.2	23.3	6.7
	500	51.5	2.1	48.3	14.2
	1000	76.7	2.7	65.7	22.9
Variance	0.5	20.7	1.9	15.4	13.9
	1	47.4	1.9	45.2	14.4
	2	71.9	2.2	76.9	15.5
Correlation	0	55.3	2.1	49.6	14.4
	0.5	38.0	1.8	42.0	14.8
Dimensionality	2	48.9	1.9	48.5	14.0
	3	45.6	2.0	44.4	15.3
Overall		46.7	2.0	45.8	14.6

Table 4

The mean of the pairwise statistics for dependent and independent item pairs as a function of the design factors

		τ		Q_3	
Design factor	Value	Dependent	Independent	Dependent	Independent
Sample size	100	0.8	-0.02	1.0	-0.36
	500	2.1	-0.04	2.0	-0.84
	1000	3.2	-0.06	3.0	-1.18
Variance	0.5	1.1	-0.02	0.8	-0.77
	1	1.9	-0.04	1.8	-0.79
	2	3.2	-0.05	3.5	-0.82
Correlation	0	2.4	-0.03	2.2	-0.78
	0.5	1.7	-0.04	1.8	-0.81
Dimensionality	2	2.1	-0.03	2.1	-0.78
	3	2.0	-0.05	1.9	-0.82
Overall		2.1	-0.04	2.0	-0.79

Table 5

The minimum percentile of the four statistics values of the (Type 1) dependent item pairs considering the distribution of all item pairs in the given cell (two-dimensional data).

Design factors			MH	G^2	τ	Q_3
Sample size	Variance	Corr.				
100	0.5	0	5.4	4.3	11.2	6.5
100	0.5	.5	5.1	4.7	3.6	5.8
100	1	0	6.9	8.7	12.0	23.6
100	1	.5	5.4	4.0	19.6	25.0
100	2	0	4.4	5.4	48.2	59.4
100	2	.5	1.8	2.5	30.4	36.6
500	0.5	0	10.1	10.1	26.1	31.5
500	0.5	.5	2.9	2.9	23.9	48.9
500	1	0	44.6	44.6	72.5	73.2
500	1	.5	37.0	35.5	67.0	77.9
500	2	0	98.6	98.9	98.9	98.9
500	2	.5	98.2	98.2	98.6	98.9
1000	0.5	0	25.4	31.5	17.4	42.8
1000	0.5	.5	9.1	8.3	56.2	60.9
1000	1	0	90.9	91.7	96.0	96.4
1000	1	.5	46.7	43.5	72.8	89.5
1000	2	0	99.3	99.3	99.3	99.3
1000	2	.5	98.6	98.6	98.9	99.3

Table 6

The minimum percentile of the four statistics values of the dependent item pairs considering the distribution of all item pairs in the given cell (three-dimensional data).

Design factors			MH		G^2		τ		Q_3	
Sample size	Variance	Corr.	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
100	0.5	0	0.7	-[b]	0.7	-	3.3	-	6.5	
100	0.5	.5	2.5	1.4	2.5	1.8	1.4	0.7	4.7	0.7
100	1	0	5.4	-	5.1	-	7.6	-	15.6	
100	1	.5	3.6	2.2	3.3	2.2	12.0	0.4	15.2	0.7
100	2	0	17.8	-	18.5	-	39.1	-	52.5	
100	2	.5	3.3	0.4	3.3	0.7	7.2	2.9	8.7	5.8
500	0.5	0	0.7	-	0.7	-	10.1	-	35.5	
500[a]	**0.5**	**.5**	**0.4**	**0.4**	**0.4**	**0.4**	**7.6**	**5.1**	**9.8**	**8.7**
500	1	0	8.0	-	8.3	-	30.4	-	37.7	
500	1	.5	3.6	1.8	2.5	1.8	30.1	6.9	49.3	8.3
500	2	0	95.3	-	96.4	-	97.5	-	97.1	
500	**2**	**.5**	**84.8**	**0.7**	**83.3**	**0.7**	**91.7**	**2.5**	**94.9**	**8.7**
1000	0.5	0	21.4	-	21.0	-	40.2	-	58.0	
1000	**0.5**	**.5**	**14.1**	**0.4**	**13.8**	**0.4**	**42.4**	**4.0**	**69.2**	**5.4**
1000	1	0	91.3	-	92.0	-	95.3	-	95.3	
1000	1	.5	71.4	3.3	66.7	2.9	84.1	13.0	91.7	14.9
1000	2	0	98.2	-	98.2	-	98.2	-	98.2	
1000	**2**	**.5**	**87.7**	**0.7**	**81.9**	**0.7**	**92.0**	**15.9**	**97.5**	**29.0**

Note a: The rows indicated in bold are used for the visual method

Note b: Type 2 is lacking when the correlation is zero

Table 7

The mean adjusted Rand index values

Design factors	Value	LOR[a]	MH	G^2	τ	Q_3
Sample size	100	0.18	0.05	0.09	0.06	0.05
	500	0.51	0.21	0.43	0.42	0.29
	1000	0.70	0.41	0.62	0.62	0.48
Variance	0.5	0.19	0.03	0.16	0.11	0.10
	1	0.44	0.17	0.38	0.33	0.23
	2	0.76	0.47	0.60	0.66	0.48
Correlation	0	0.47	0.31	0.42	0.40	0.33
	0.5	0.46	0.14	0.35	0.33	0.21
Dimensionality	2	0.50	0.26	0.39	0.39	0.29
	3	0.43	0.19	0.37	0.34	0.26
Overall		0.46	0.22	0.38	0.36	0.27

Note a: LOR denotes log-odds ratio.

Table 8

The percent of correct decisions on the number of item properties as a function of the design factors

Design factors	Value	LOR[a]	MH	G^2	τ	Q_3
Sample size	100	40.8	35.0	56.4	50.0	34.0
	500	51.7	36.7	48.3	48.3	53.3
	1000	57.5	36.7	40.8	65.0	60.8
Variance	0.5	43.3	41.6	52.5	40.8	42.5
	1	52.5	37.5	45.8	51.7	36.7
	2	54.2	29.2	43.8	72.3	69.1
Correlation	0	56.1	35.6	45.8	59.3	57.8
	0.5	43.9	38.7	45.1	49.7	41.1
Dimensionality	2	67.2	45.6	57.3	68.6	61.7
	3	37.8	26.7	37.2	41.0	37.2
Overall		50.0	36.2	45.4	54.6	49.4

Note a LOR denotes log-odds ratio.

Figure captions

Figure 1

Theoretical grey-scale matrices for the four selected design cells

Figure 2

Grey-scale matrices for three-dimensional data with N=500, variance 0.5, correlated dimensions

Figure 3

Grey-scale matrices for three-dimensional data with N=500, variance 2, correlated dimensions

Figure 4

Grey-scale matrices for three-dimensional data with N=1000, variance 0.5, correlated dimensions

Figure 5

Grey-scale matrices for three-dimensional data with N=1000, variance 2, correlated dimensions

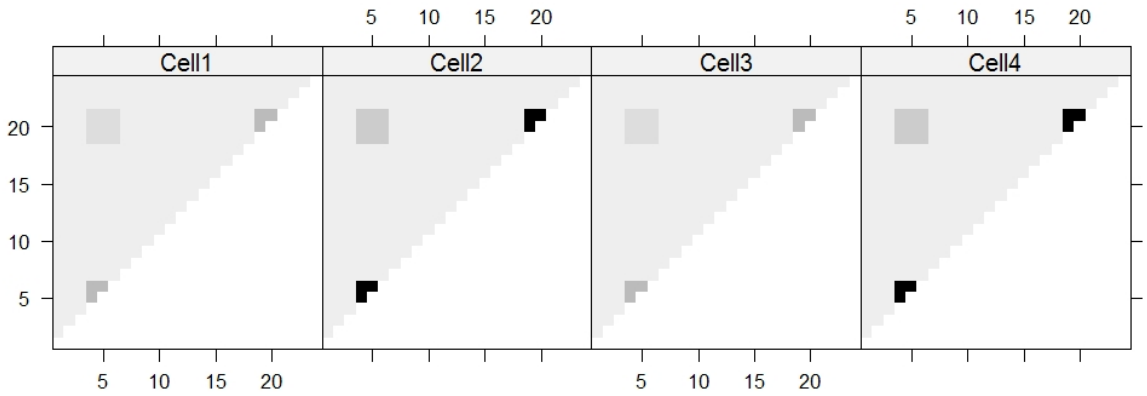


Figure 1

Theoretical grey-scale matrices for the four selected design cells. All cells refer to three dimensional data with correlated dimensions (r=.5). The data sets corresponding to Cell 1 are with N=500, variance of 0.5; to Cell 2 with N=500, variance of 2; to Cell 3 with N=1000, variance of 0.5; and to Cell 4 with N=1000, and a variance of 2.

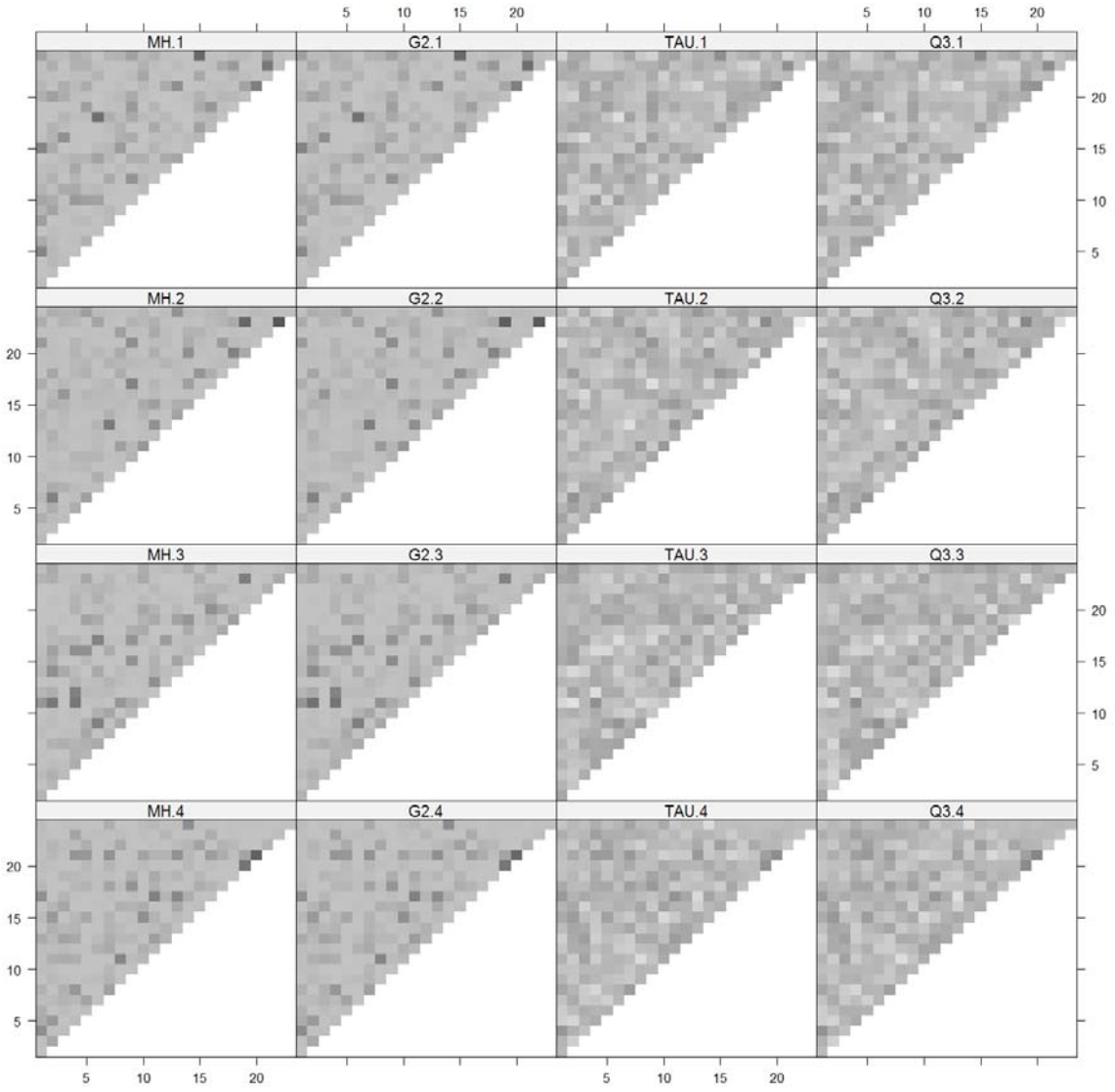


Figure 2

Grey-scale matrices for three-dimensional data with N=500, variance 0.5, correlated dimensions (four data sets)

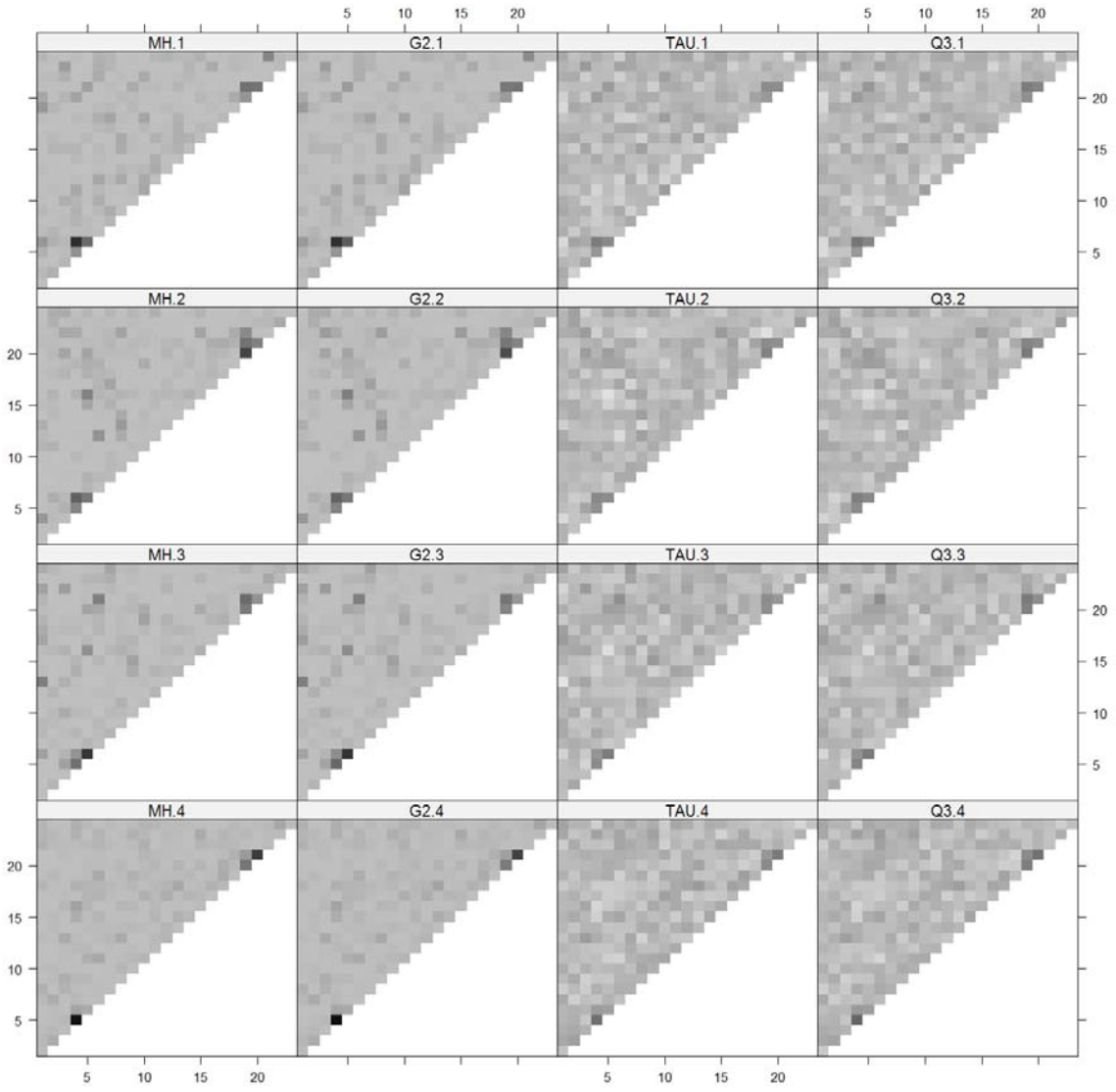


Figure 3

Grey-scale matrices for three-dimensional data with N=500, variance 2, correlated dimensions

(four data sets)

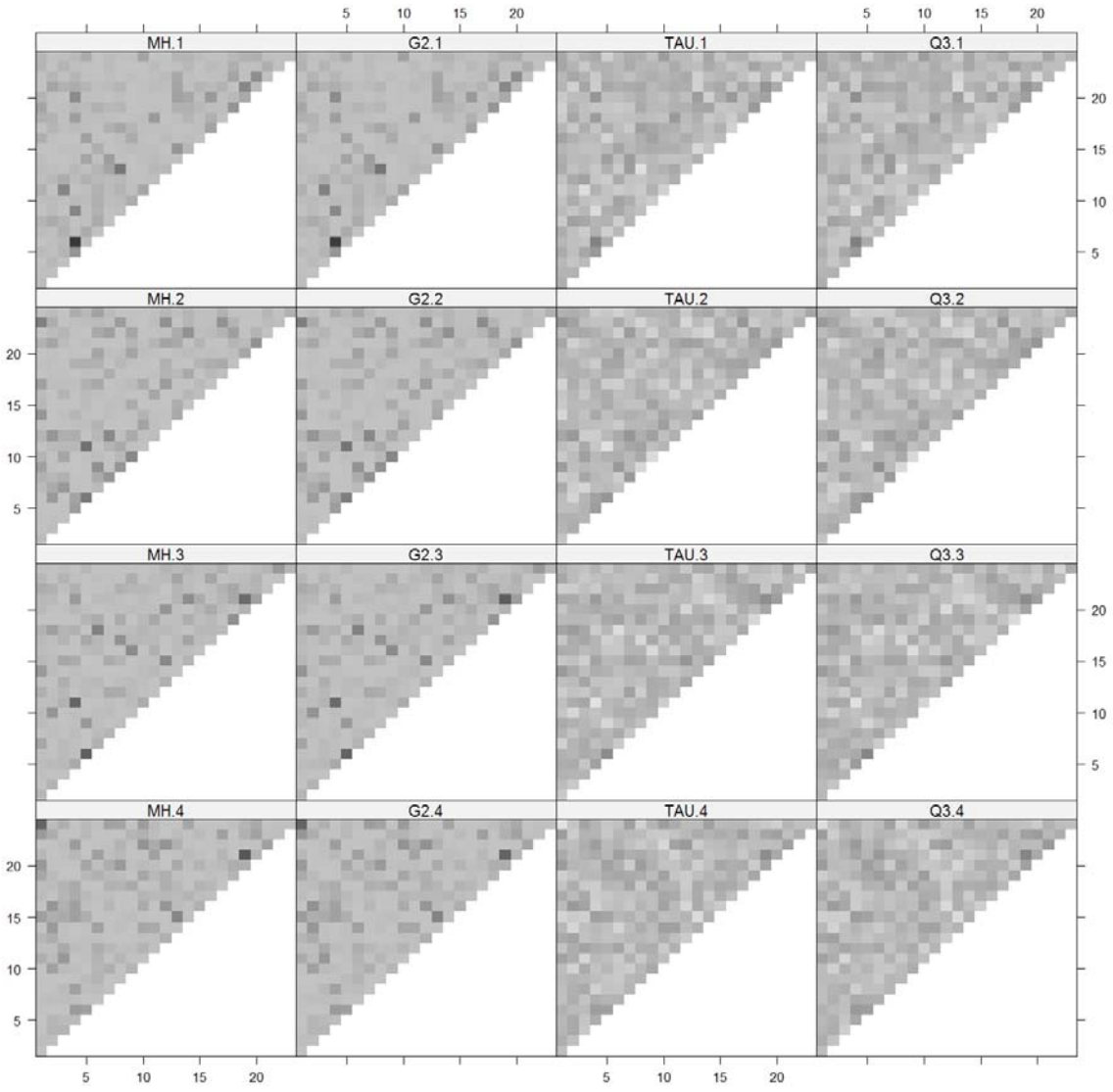


Figure 4

Grey-scale matrices for three-dimensional data with N=1000, variance 0.5, correlated dimensions (four data sets)

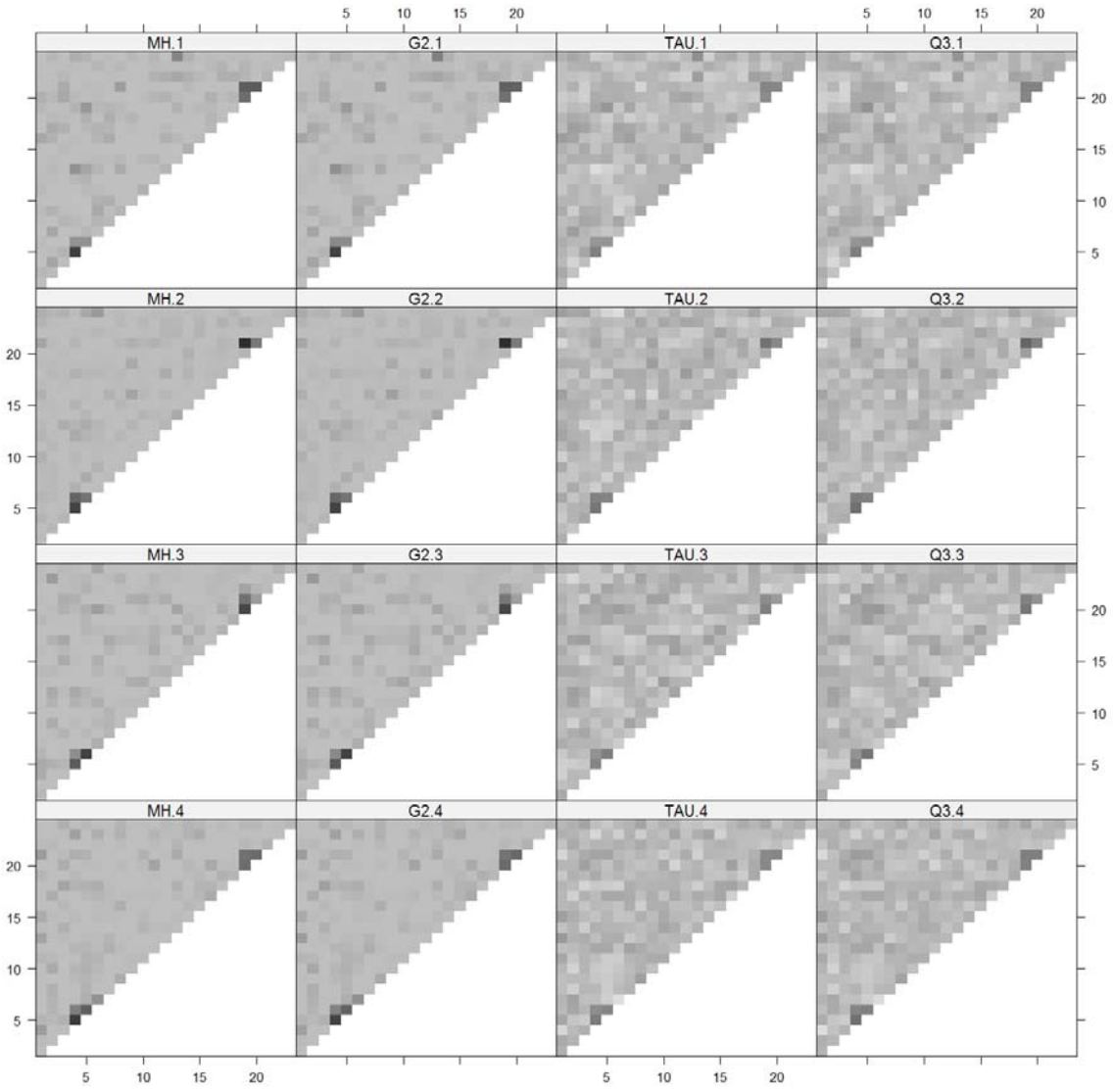


Figure 5

Grey-scale matrices for three-dimensional data with N=1000, variance 2, correlated dimensions (four data sets)