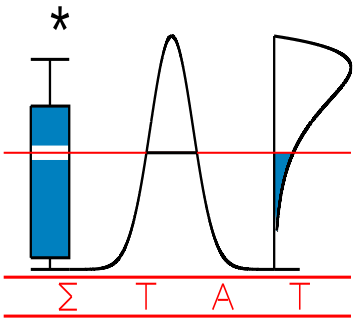
T E C H N I C A L

R E P O R T

0685

THE DETECTION OF HIDDEN ITEM PROPERTIES AND MULTIDIMENSIONALITY

BALAZS, K., SCHEPERS, J. 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: DETECTION OF HIDDEN ITEM PROPERTIES

The detection of hidden item properties and multidimensionality

Katalin Balázs, Jan Schepers and Paul De Boeck

K.U. Leuven, Belgium

The detection of hidden item properties and multidimensionality

Abstract

It is not uncommon in psychometric modeling to consider the effects of item properties on the item responses, for example to explain item difficulties. An important reason is that knowing these effects contributes to a better understanding of the responses, and hence of what is being measured. The effects of item properties may be different depending on the person, and therefore the item properties can be a source of multidimensionality. When the item properties are known preliminary to the analyses, methods such as the random-weights LLTM, and alternating logistic regression (ALR), a GEE based method, can be used in order to take the individual differences into account. Here, a method is investigated for the case when there is no advance knowledge on the item properties, and when the properties are to be detected from the data instead (latent item properties). The method is based on an (1) additive clustering algorithm (ADCLUS) to extract binary properties from the pairwise association values, and (2) alternating logistic regression (ALR) to model the effects of the extracted item properties.

In a first simulation study, the combination of ADCLUS for property extraction and ALR for modeling the property effects was investigated. The results show that ADCLUS can reveal the latent item properties, but also that it is difficult to decide on the exact number of latent item properties. The statistical significance of the ALR modeled effects of the ADCLUS extracted properties did not turn out to be a good criterion, but with the ADCLUS explained variance or with the ALR weights as a criterion much better results were obtained.

In a second simulation study with three different data structures, the performance of ADCLUS is compared to the performance of DETECT, a procedure for binary data that is developed to detect multidimensionality and to reveal item partitions. A combination of the two procedures is also considered. When a selection needs to be made, the combined method seems the most appropriate choice. However, when the item clusters are clearly overlapping, ADCLUS seems a better choice.

Depending on the domain in the social sciences where questionnaires or tests are used, a different importance is given to item properties. Item properties may evoke abilities and attitudes and play an important role in the response process and individual differences therein. The consideration of these aspects can provide valuable information on the underlying structure of the data and can lead to a deeper understanding of a phenomenon under study.

The item properties may be known a priori to the analyses, when the items have an apparent structure or when the items are designed to have certain properties. For example, when a set of item properties is systematically manipulated in a test, the test has an apparent design (Embretson, 1985). In cognitive psychology, for example, a carefully controlled test design is not uncommon for the study of intelligence (Sternberg, 1977a), deductive reasoning (Sternberg, 1977b, 1979; Rijmen & De Boeck, 2002), analogies (Whitely, 1978), and spatial representations (Egan, 1979; Embretson, 1985, chap. 3). The idea of item properties is the basis of the cognitive diagnostic approach (Tatsuoka & Tatsuoka, 1982), which represents the item properties in the so-called Q-matrix (Tatsuoka, 1990). The difference with the previous approaches is that the Q matrix does not necessarily stem from a test design, but often is the result of expert judgments.

MIRT models such as the Multicomponent Latent Trait Model and the General Component Latent Trait Model formulated by Embretson (1980, 1984), the fusion model (DiBello, Stout, & Roussos, 1995; Hartz, 2002), the conjunctive Rasch model (Maris, 1995), the DINA (Junker & Sijtsma, 2001) and NIDA models (Junker, 2001; Junker & Sijtsma, 2001), the random weights LLTM (Rijmen & De Boeck, 2002), alternating logistic regression (Hardin & Hilbe, 2003) and regular confirmatory versions of the multidimensional 2PL model (McKinley, 1989), all are methods which can be applied to a binary data matrix starting from a given item property matrix or Q matrix.

Person based random effects of the item properties or mastery probabilities based on these properties (as in the cognitive diagnostic approach), do explain individual differences in performance on the item properties. For instance, some may have problems with reasoning tasks because their performance is affected more than the performance of others by negations in the premises. Similarly, some examinees are very sensitive to the evaluative connotation of items in an attitude questionare, while others are much less sensitive. Individual differences based on the effects of item properties would lead to associations between items sharing the same property.

As it has been mentioned, the potential of person-based effects of item properties is not recognized equally well in all fields of assessment. Nevertheless, researchers in domains such as personality psychology and cognitive psychology are often interested precisely in such individual differences because they define dimensions of individual differences.

In personality psychology, the study of individual differences based on situational features is called *interactionism* (e.g., Blumer, 1969; Pervin, 1977). In the domain of intelligence and cognitive tasks in general called the *cognitive components approach*. Following this approach, it is investigated how much a certain process (induced by an item property) contributes to the probability of success or to the response time (Embretson, 1985; Sternberg , 1977a). In the latter case, the weights of the item properties in the probability function or the response time function are assumed to show individual differences, referring to differences in accuracy or speed, respectively, associated with the processes that correspond to the item properties.

In the previous approaches, the item properties are given and are therefore available for analysis of the data. However, the item properties can as well be unknown, so that explorative techniques would be required to reveal the underlying properties. The present study is planned to address this issue for binary data and binary latent item properties. The

aim is to reveal item properties which evoke processes that differentiate between persons and therefore are at the basis of the dimensions in the data structure. Consequently, the investigated problem can be considered as a special type of dimensionality analysis where the dimensions are linked to binary item properties. An approach will be presented which is based on a combination of additive clustering with marginal modeling (with GEE). Additive clustering is used in the first step, to extract item properties, and marginal modeling is used in the second step, in order to test the effects of the extracted item properties on the associations (as in the alternating regression approach, ALR; Hardin & Hilbe, 2003).

First, it will be explained how one can proceed when the item properties are given, and next, a method will be presented to extract the item properties from the data when they are not given, but hidden instead.

Approach for known item properties: GEE with ALR

When binary responses are collected from a test with known (manifest) item covariates, the random weights linear logistic test model (RWLLTM; Rijmen & De Boeck, 2002) is an evident way within item response theory to model the random effects of a given set of item covariates:

$$\text{logit}\left(P\left(Y_{pi}=1\middle|\theta_p,\boldsymbol{\beta}_p\right)\right)=\theta_p+\sum_k \beta_{pk}X_{ik}\,, \qquad (1)$$

where $P\left(Y_{pi}=1\middle|\theta_p,\boldsymbol{\beta}_p\right)$ is the success probability of person p for item i, modeled as a function of covariates. X_{ki} is the value of the k-th covariate (k=1,…, K) for item i (i=1,..., I), and the β_{pk} is the associated random weight for person p. θ_p is the random intercept that is the so-called ability of the person in the context of achievement tests. When the latter is the only source of individual differences, the resulting model is the Linear Logistic Test Model

(LLTM; Fischer, 1973). When, additionally, $K=I$, and the K vectors form an identity matrix, the equation yields the Rasch model (1960). Individual differences in the effects of item covariates are indicated with the subscript p for β, as in β_{pk}, and a multivariate normal distribution is assumed for $(\theta_p, \boldsymbol{\beta}_p)$.

This approach is rather cumbersome when the number of item covariates is high and many combinations need to be tested to find out whether individuals differ with respect to the covariate effects. The estimation of models with a large number of dimensions involves a high number of quadrature points when the multidimensional integral is approached with a method such as the Gaus–Hermite method. One may consider a Bayesian approach instead (e.g., Bèguin & Glas, 2001; Segall, 2001), but the many combinations of item covariates would require a large number of analyses nevertheless.

An alternative way of tackling this problem is a marginal modeling approach for the means and associations based on all item properties. Because with ALR (Hardin & Hilbe, 2003) models with high dimensionality can be tested very easily, high numbers of item properties are not a problem. Individual parameter estimates for the persons cannot be obtained, but once the properties are identified which show individual differences, in a final step a RWLLTM model can be fitted, so that individual estimates are obtained. Here, the focus is not on measuring abilities but on identifying properties with an effect that differs between individuals.

Generalized estimating equations (GEE; Liang & Zeger, 1986) form a specific group of marginal models. GEE models were developed for extending the Generalized Linear Models to correlated data. The term itself indicates that these estimating equations are not based on the likelihood, but that they are generalizations of other estimating equations instead (Hardin & Hilbe, 2003). Therefore, GEE approaches are less computationally intensive than the full likelihood approach. The effects of the covariates and the association parameters are

much simpler to estimate computationally with the GEE approach than with the full information likelihood, because GEE uses only part of the information of the data, such as the means, marginal correlations, and constrained marginal association parameters for binary data (Liang, Zeger, & Qaqish, 1992; Ziegler, Kastner & Blettner, 1998), so that there is no need to specify the full likelihood of the data.

On the other hand, GEE models have drawbacks, which actually follow from their advantages compared to likelihood based methods. GEE does not provide the traditional fit statistics, although fit statistics have been developed (e.g., Barnhart & Williamson, 1998; Pan, 2000; Zheng, 2000) for GEE1, which focuses on the first moment.

Several extensions of the original GEE were developed for modeling the associations. In GEE1, the variables are treated as if they were orthogonal by way of working assumption (Liang, Zeger & Qaqish, 1992). The quality of the working assumptions on the association parameters does not detract from the consistency of the estimates, but the working assumptions can reduce the preciseness of the parameter estimates (Prentice & Zhao, 1991).

In the GEE2 variant (Prentice & Zhao, 1991; Zhao & Prentice, 1990), the mean and the association structure are jointly estimated, but the correctness of the assumptions to be made in order to estimate the structure, does affect the estimates (Hardin & Hilbe, 2003). The estimated variance of the parameters may not always be consistent when the associations are misspecified. Furthermore, GEE2 cannot handle very many items.

Thus far, for our purpose, Alternating Logistic Regression (ALR; Carey, Zeger, & Diggle, 1993; Liang, Zeger, & Qaqish, 1992) seems the most appropriate GEE variant. Alternating logistic regression combines a marginal logistic regression for the mean structure with a logistic regression for the association structure. The name of the method stems from the fact that the estimation process alternates between these two logistic regression steps: (1) the parameters of the mean structure, the βs (see Equation 2), are estimated in a marginal

logistic regression using a first order GEE, and the current association parameter estimates and (2) the parameters of the association structure, the αs (see Equation 5), are estimated using a logistic regression of y_{pi} on each $y_{pi'}$ ($i>i'$), and the current estimates of the mean structure parameters.

In this way, unlike in a regular GEE1, the association structure (second moment) is estimated. ALR requires association covariates but these can be derived easily from the single item covariates. Each single item covariate defines also an association covariate by taking the product of the single item covariate values, as in Equation 4. This procedure can handle a rather large number of items and it is implemented in many software packages (Hardin & Hilbe, 2003).

In the first component of the ALR approach, the marginal odds are modeled as in

$$\text{logit}(P(Y_{pi})) = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} \qquad (2)$$

where β_0 is the overall difficulty parameter, X_{ki} is the *k*-th item covariate changing its value over items (*i*), and β_k is the corresponding weight.

Because the αs from Equation 5 are included in the GEE equations for β, the current estimates for the αs are used.

In the second component of the ALR approach, marginal odds ratios are used as the measure of association:

$$OR(Y_{pi}Y_{pi'}) = \frac{P(Y_{pi}=1,Y_{pi'}=1)P(Y_{pi}=0,Y_{pi'}=0)}{P(Y_{pi}=1,Y_{pi'}=0)P(Y_{pi}=0,Y_{pi'}=1)}\ , \qquad (3)$$

where *p* refers to a person, *i* and *i'* are items of the same item pair.

A logistic regression model is fitted to obtain an estimation of the covariate effects on the log-odds ratios:

$$\log\left(OR\left(Y_{pi}Y_{pi'}\right)\right) = \alpha_0 + \sum_{k=1}^{K} \alpha_k X_{ki} X_{ki'}, \qquad (4)$$

where X_{ki} and $X_{ki'}$ are the values of items *i* and *i'* on the *k*-th binary item covariates, and α_k is the association parameter belonging to the *k*-th item covariate, in other words α_k is a weight that indicates how much item covariate k contributes to the log-odds ratio, and α_0 is an overall association parameter for all the item pairs. The pairwise products of the binary item covariates define the so-called association covariates. Using these association covariates, $Z_{kii'} = X_{ki}X_{ki'}$, yields Equation 4.

$$\log\left(OR\left(Y_{pi}Y_{pi'}\right)\right) = \alpha_0 + \sum_{k=1}^{K} \alpha_k Z_{kii'} \qquad (5)$$

Because the βs from Equation 2 are included in the GEE for the estimation of the αs, the current values of the βs are used.

The parameter α_0 represents a common underlying source of individual differences. For the Rasch model and the LLTM, only α_0 would be different from zero. The K association covariates explain the association beyond this common source, so that the association covariates define possibly overlapping item clusters with an additional association, and therefore also an additional dimension. Given the random effects model the weight parameter

α_k expresses the weight of the random effects β_{pk} (see Equation 1) in explaining the log-odds ratio.

The ALR variant of marginal modeling has been shown effective in detecting sources of person based heterogeneity in data with a RWLLTM type of true structure (Balazs, Hidegkuti, & De Boeck, 2006). The ALR approach is an attractive one because it is not time consuming and still gives stable and straightforward parameter estimations. Furthermore, it can be used as a basis for a full likelihood based analysis, where a random effects model is built in line with the obtained ALR results.

How to extract unknown item properties: ADCLUS

Even when a test is not constructed based on item properties, there might still be hidden or latent item properties with random effects and thus item properties functioning as sources of individual differences, each defining a different dimension. The detection of latent item properties with random effects is equivalent to the detection of item clusters. Note that the item properties may overlap, so that also the item clusters would overlap. In Study one, a method for overlapping clusters is used, called ADCLUS. Also DETECT (Stout, Habing, Douglas, Kim, Roussos & Zhang, 1996) aims at detecting dimensionality and item clusters, but DETECT requires a simple structure. A simple structure corresponds to non-overlapping clusters. DETECT will be investigated in comparison with ADCLUS in Study two.

In order to be consistent with the previously described ALR approach, the clustering method for overlapping clusters should (1) concentrate on the marginal log-odds ratios such as ALR does, and (2) be based on an additive model. An additive clustering method called ADCLUS (see e.g., Arabie & Carroll 1980; Carroll & Arabie, 1983; Lee, 1999, 2001;

Shepard & Arabie, 1979) fulfills these two criteria when applied to the log-odds ratios as similarities. In the ADCLUS model, the similarity of two objects is described in the following equation:

$$s_{ii'} = w_0 + \sum_{k=1}^{K} w_k \xi_{ki} \xi_{ki'}, \qquad (6)$$

where $s_{ii'}$ is the similarity between items *i* and *i'*. ξ_{ki}, is an underlying binary item property, so that its value is one if item *i* has property *k*, and zero otherwise; w_k, the weight of product $\xi_{ki}\xi_{ki'}$, denotes the importance of property *k* in the similarity, and w_0 is an additive constant. In Lee's algorithm, the ADCLUS algorithm which will be focused on, the weights have always nonnegative values.

It is assumed in the proposed approach that the shared properties (common properties) contribute in an additive way to the similarity, and that the shared absence of properties does not contribute. Note that, except for the latent nature of the ξ s, when log-odds ratios are used as similarities, Equation 6 is formally equivalent to Equations 4 and 5. The equivalence of Equations 4, 5 and 6 enables us to use the properties obtained by additive clustering in an ALR model. Accordingly, a first step with ADCLUS can provide the required covariates for ALR in a second step. Applying ALR on the ADCLUS extracted properties provides an interesting statistical tool to investigate the effects of latent binary item properties, or in other words, the effective part of a Q matrix that is extracted from the data. One should be aware, however, that this procedure is vulnerable to the capitalization on chance, because the properties are extracted from the data, and next, they are tested using the same data. Therefore, we will also use a cross-validation method, with half of the data being used for the property extraction and the other half for significance testing in ALR. Furthermore, in ADCLUS, the dependence of the association parameters on the parameters of the mean

structure is not taken into account, so that the ADCLUS analysis must be considered as an approximate method.

The ADCLUS algorithm

Shepard and Arabie (1979) have proposed the original additive clustering algorithm, called ADCLUS. Based on their approach, several alternative algorithms have been developed (e.g., Arabie & Carroll, 1980; Carroll & Arabie, 1983; Chaturvedi & Carroll, 1994; Hojo, 1983).

Mirkin (1987, 1989) suggested a general algorithm, in which the clusters are extracted sequentially, as follows. After each extraction step, the weights of the item properties are recalculated including the newly extracted cluster, and the residual similarity matrix is determined in preparation of the extraction of the next cluster. From various methods suggested by Mirkin for the derivation of the clusters, the ADDI-S algorithm is the most prominent one (Mirkin, 1987).

In ADDI-S algorithm, the cluster extraction begins with the selection of the pair of items with the highest similarity. After selecting the first item pair, the algorithm selects from the remaining items the item with the largest average similarity to the selected items. The expansion of a cluster terminates based on a criterion using the internal and external similarity measures. The internal similarity is the average similarity within the cluster, the external similarity is the average similarity between elements not belonging to a cluster. When an item has an average similarity with the cluster items that is larger than halfway the difference between the external and internal similarity, than the item is included in the cluster. If none of the item fulfills this criterion, then the expansion of the cluster terminates. After the

extraction, the residual similarities are calculated, and the process of extracting a next cluster is started based on the same principles.

As a new development, Lee (2001) proposed an algorithm where the BIC measure is used to define model complexity of additive clustering models. Lee's approach offers an interesting kind of control over the trade-off between model preciseness and complexity. Lee (2001) uses a formulation of the maximum likelihood for additive clustering as provided by Tenenbaum (1996), that is, the probability of a similarity matrix given the derived cluster structure and the corresponding weights. The proposed BIC formula is presented in the following.

$$BIC_{ADCLUS} = \frac{1}{pre^2} \sum_{1 \le i < i' \le I} (s_{ii'} - \hat{s}_{ii'})^2 + K \log\left(\frac{I(I-1)}{2}\right), \qquad (7)$$

where I is the number of items, K is the number of clusters and thus K+1 is the number of parameters in the model (counting also w_o), $s_{ii'}$ is the observed and $\hat{s}_{ii'}$ is the estimated similarity between items i and i'. The proposed BIC measure is basically the sum of squared errors (SSE), scaled by the precision (*pre*) (see e.g., Tenenbaum, 1996). The precision (*pre*) is defined as follows:

$$pre = \frac{1}{I(I-1)/2} \sum_{1 \le i < i' \le I} \sqrt{\frac{\sum_{p=1}^{P} \left(s_{ii'}^{p} - \bar{s}_{ii'}\right)^2}{P-1}}, \qquad (8)$$

where P is the number of persons, $s_{ii'}^{p}$ and $\bar{s}_{ii'}$ are the similarities of item i and item i' for person p, and on average, respectively. Note that lower precision values mean smaller variance of across similarity matrices and hence a higher precision, as can be seen in Equation 8. A condition for the precision number to be determined, is that similarity values for different entries (persons, groups) are available. In practice, the similarities often stem from direct ratings by individual persons.

Lee (2001) applied an additive clustering algorithm with the BIC measure as a criterion and consisting of five steps. Steps from Step 2 to 4 are repeated several times.

Step 1: The BIC measure is calculated for a model with only a general cluster. By definition, the model in Step 1 always explains zero percentage of the variance of the similarities.

Step 2: A new cluster is generated based on Mirkin's (1987) ADDI-S algorithm.

Step 3: The aim of this step is to avoid local minima. It starts with a combinatory part based on stochastic hill-climbing. Starting from a random ordering of the elements of the cluster membership vector, the value of each element in turn is changed from zero to one or from one to zero, depending on the original value. After each change, the property weights and the additive constant are derived by a non-negative least-square optimization (Lawson & Hanson, 1974), and the sum of squared error (SSE) is calculated. If the SSE decreases by changing the cluster membership, the new cluster structure is accepted, and a new random ordering of the cluster memberships is generated and the membership change starts following the new order. Otherwise, the original cluster membership is kept and the next membership value is changed. When all cluster memberships changes has been tried out, three cluster memberships are replaced randomly, and the hill-climbing restarts. The hill-climbing process restarts till this change of the randomly selected three items is carried out as many times as the predefined patience number is without any improvement during the process. In the studies described in the following, the patience number was set to 50.

Step 4: If the BIC of the new model is smaller than that of the previous model with at least the value of the evidence parameter, the program calculates the residual similarities and returns to Step 2, and an additional cluster is added, otherwise it continues with Step 5. Kass and Raftery (1995) proposed a scale for the BIC differences which is called the *evidence*

parameter. Lee's algorithm uses an evidence parameter of 6 as a default stopping rule, but it can be changed by the user.

Step 5: The algorithm terminates. The additive clustering model with the lowest BIC is selected, and the cluster memberships and the corresponding weights are shown.

A similar ADCLUS algorithm is also available for a prespecified number of clusters. The algorithm with a prespecified number of clusters is much simpler. First, a random cluster membership matrix is generated with as many clusters as the predefined number of clusters. Afterwards, the weights of the clusters are calculated, and different from the earlier described algorithm, before the hill-climbing, three randomly selected elements of the least weighted cluster are changed. Next, the hill-climbing starts as it has been described for the other algorithm in Step 3. In this algorithm, the overall process terminates when the procedure have changed three randomly selected elements of the least weighted cluster without an improvement as many times as the patience number is.

Because it was observed that the ADCLUS algorithm with fixed cardinality sometimes terminates at a local minimum, a small change of the algorithm has been implemented. After trying out the change of each element of the cluster matrix, instead of changing three elements of the least weighted cluster, a new cluster membership matrix is generated. According to our observations, at a given patience number, the modified algorithm performs better. In the simulation studies described in the following, when a pre value is not calculated, the modified algorithm is used for item property extraction.

An additional complication is that in our application there is not a set of similarities available for each person, because the log-odds ratios cannot be determined for each individual separately. Hence the precision number cannot be determined. In such cases, an approximation of the precision can be used based on former experience. Lee (2002, p. 43)

suggests to apply *pre* values of 0.05, 0.1, 0.15 with normalized similarity matrices, corresponding to 'precise', 'average' and 'imprecise' data sets, respectively.

Two simulation studies are planned. In Study one, the performance of ADCLUS in combination with ALR is studied, as an approach to extract and test hidden item properties. In Study two, ADCLUS and DETECT (an algorithm for detecting partitions) are compared in revealing latent overlapping and non-overlapping structures, although DETECT is meant only for the latter. However in practice one does not have advanced knowledge of the underlying structure before the analysis, hence, it is interesting to see how the two methods perform for both structures. These two simulation studies are described in detail in the following sections.

Study one

The performance of ADCLUS was studied for binary data with an underlying overall random effect and two additional random effects related to the item properties. The design was planned in agreement with common data set features in psychological applications: (1) Often the sample size is rather small, as small as 100 or several hundreds. (2) Often the number of items is rather small, as in single-trait or single-attitude scales. (3) Often there is a common dimension, and in addition a few, possibly less important dimensions. Besides, several challenging conditions were also looked at, such as highly correlated dimensions, dimensions based on a relatively small number of items, and dimensions which overlap (lack of simple structure).

The data were generated based on the following RWLLTM structure:

$$\text{logit}\left(P\left(Y_{pi}=1\middle|\theta_p,\boldsymbol{\beta}_p\right)\right)=\theta_p+\beta_{1p}x_{1i}+\beta_{2p}x_{2i}+\beta_i, \qquad (9)$$

The structure as defined in (9) is one with a general dimension and two, possibly overlapping, more specific dimensions, to be described further. In Study one, all data sets were generated with 24 items and with two possibly overlapping item properties. The sample size was 100, 250, 500, 750 or 1000. β_i, the item difficulty parameter was generated with a normal distribution (over the items) with a mean of zero and a variance of one. The probability that a property is present (has a value of 1 instead of 0) in the item was .3, .5 or .7 for both properties, yielding smaller or larger item clusters. The properties were generated independently. The random weights of the properties (weights are random over persons) had a normal distribution with zero mean and a variance of either 1 or 2 for both properties. The correlation among the three random effects was either 0 or .5. Given that one of the three random effects (the random intercept) contributes to all items, there is always a general dimension present in the data. This issue will be investigated in a following section. The overall success probability was .5. The general ability of person p, θ_p, had a normal distribution with zero mean and a variance of one. The three probability values of the properties, the two variance values for the random effects of the properties, the two correlation values among the random effects, and the five sample sizes were fully crossed in the design. Ten data sets were generated per design cell, so that 600 data sets in total were generated.

The two research issues are: (1) can the true properties be recovered from the data given that the true number of properties is known, and (2) can the true number of properties be determined from the data. For all data sets, two is the true number of item properties.

First, the performance of ADCLUS was studied, in order to find out whether it can reveal the true item properties given that the number of true properties is known. In addition to the additive constant, two properties were estimated for each data set. 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. In our application, the two properties define a partition corresponding to four patterns: 11, 10, 01, 00. Steinley (2004) has shown in a simulation study that the correct classification rate is mostly higher than the adjusted Rand index value. He also provided a rough guide to assess the quality of partition agreement: an adjusted Rand index value greater than 0.9 is excellent, between 0.8 and 0.9 is good, between 0.65 and 0.8 is moderate, and smaller than 0.65 is poor.

Second, it was studied whether the correct number of item properties can be found from the ALR analyses when a larger number than the true number of properties is extracted through ADCLUS. The ALR method is considered to be the best candidate for modeling the effect of the extracted item properties, because the model formulation of ALR is equivalent to that of ADCLUS, and because ALR is a very convenient method for modeling associations. In addition, also ADCLUS itself was investigated on its potential to reveal the true number of clusters.

Results

Recovery of the true properties

For all data sets, pairwise log-odds ratios were calculated. The log-odds ratios were used as input to ADCLUS with a prespecified number of two clusters.

When the extracted item properties are compared to the true ones, it is not always evident which combination one should use. Therefore, the adjusted Rand index was calculated for both possible combinations and the higher value was chosen of the two. The overall mean of the adjusted Rand index values comparing the extracted properties to the true ones was

0.72, the variance was 0.08. The obtained adjusted Rand index values have a negatively skewed distribution. Considering all data sets, and following Steinley' s (2004) categorization, the recovery was excellent for 41.5% of the data sets, good for 9.3%, moderate for 12.5% and poor for 36.7%. However, the performance is clearly a function of the design factors as it will be demonstrated in the following.

Based on a linear regression analysis, the sample size has significant linear, quadratic and cubic effects on the adjusted Rand index values. Figure 1 shows the effect of the sample size. The average adjusted Rand index values for samples of 100, 250, 500, 750, 1000 are 0.38, 0.66, 0.81, 0.88, and 0.89, respectively, meaning that from a sample size of 750 on, an almost excellent average recovery is obtained. Based on the results, a sample size of at least 500 is to be recommended in order to obtain a good recovery. Hence, for further analyses in this section, only the data sets with a sample size of at least 500 are considered. For these data, the average adjusted Rand index value was 0.86, and the variance of the adjusted Rand index values was 0.05. Taking into account only the data sets with a sample size of 500 or higher, the recovery was excellent for 61.7%, good for 8.9%, moderate for 12.8%, and poor for 16.7% of the data. Given that the design contains some rather challenging cases (e.g., a correlation of 0.5 and a property probability of 0.3, implying correlated and small clusters), it can be considered a very good performance. Note that even when the correlation between the two random effects is zero, there is still a positive correlation between all items because of the random intercept or general dimension (see Equation 9).

Insert Figure 1 about here.

From a multiple regression with all main effects, pairwise and triple interactions (for sample sizes from 500 on), it turns out that all main effects are significant with $p<.05$. The probability of the item property, the sample size and the variance of the random effects all have positive effects, while the correlation of the random effects has a negative effect on the adjusted Rand index. In addition, the following five interactions were also significant ($p<.05$): (1) the interaction of the sample size and the correlation of the random effects ($p<.001$), (2) the interaction of the dimensional variance and the sample size ($p<.01$), (3) the interaction of the dimensional variance and the probability of the item properties ($p<.001$), (4) the interaction of the dimensional variance and the correlation of the random effects ($p<.05$), (5) the interaction of the probability of the item properties, the correlation of the random effects and the sample size ($p<.05$). The results are described further in the following.

From Figure 1 to 4, the results are presented in box plots. The dark boxes represent the middle half of the data, the white lines in the black boxes mark the median. The lines outside the boxes marking an interval, show the upper and lower quartiles excluding the outliers. The outliers are represented by separate lines outside the interval. It is clear from Figure 1 that after omitting the data sets with small sample sizes (100 and 250), the sample size still has an effect, but a less strong one than before. Note that in Figures 2 to 4, the results relate to sample sizes of at least 500.

It is not surprising that the more items share an item property, the better the recovery is. Accordingly, it can be seen in Figure 2 that a higher probability of the item properties leads to higher adjusted Rand index values. A probability of .5 yields an equal probability distribution of item property patterns (.25 for each of the four patterns 11, 10, 01, 00). The probabilities of .3, .5, and .7 yield average adjusted Rand index values of 0.77, 0.87, and 0.95, respectively.

Insert Figure 2 about here.

The dimensional variance also has a positive effect on the Rand index values. As it is shown in Figure 3, the larger the variances of the random effects linked to the item properties are, the more effective ADCLUS is in revealing the true properties. The average recovery is good (0.81) for a variance of 1, and it is excellent (0.92) for a variance of 2.

Insert Figure 3 about here.

In general, the correlation of the random effects reduces the performance of ADCLUS, as it can be seen in Figure 4. The average adjusted Rand index is 0.94 for uncorrelated random effects and 0.78 for correlated random effects (with a correlation of 0.5).

Insert Figure 4 about here.

In the following, the significant interactions are discussed. Also these Figures are based on a sample size of at least 500. A larger sample size leads to a better recovery, but the effect is slightly different depending on the correlation, as it is shown in Figure 5. The average adjusted Rand index values for data with a sample size of 500, 750 and 1000 are 0.75, 0.78 and 0.82 in case of correlated random effects and 0.88, 0.97, 0.97 in case of uncorrelated

random effects, respectively. When the correlation is zero, a ceiling value is reached from N= 750 on.

Insert Figure 5 about here.

The interaction of the dimensional variance and the sample size is of a kind that the sample size has a stronger effect when the variance is small, as is shown in Figure 6, but the smaller variance is not fully compensated by the sample size up to a level that is reached for a large variance.

Insert Figure 6 about here.

Furthermore, there is also an interaction of the probability of the item properties with the variance of the random effects (Figure 7). The smaller the variance is, the stronger the effect of the probability is. In this case, the probability does (almost) compensate for a small variance. The combination of .7 as a probability and two as a variance yields an average adjusted Rand index of 0.97, and when the variance is only one and the probability is still .7, an average adjusted Rand index of .94 can be reached.

Insert Figure 7 about here.

Finally, also the correlation interacts with the variance, as shown in Figure 8. The correlation has a stronger effect for a smaller variance. For a correlation of zero and .5, when the dimensional variance is one, the average adjusted Rand index values are 0.90 and 0.71, respectively, while when the variance is two, the corresponding Rand index values are 0.98 and 0.85, respectively. It is a general phenomenon for the pairwise interactions that the effect of a factor is larger for the level of the other factor with a lower performance.

Insert Figure 8 about here.

Finally, also a triple interaction was found, which can be interpreted in a similar way. When the correlation is high and the sample size is small, the effect of the probability of the item properties is stronger.

In sum, given that the true number of item properties is known, ADCLUS reveals the underlying item properties rather well if certain conditions are met. In general, a sample size of 500 seems to be sufficiently high to effectively reveal the unknown item properties, especially for a large variance (2) or a high probability of the item properties (from .5 on). The design factors interact in such way that the effect of a factor is larger for the level of the other factor with the poorest performance, but often a full compensation is not possible.

Correlation between the item clusters

The correlation of the two random effects that are linked to the item properties were 0 or .5 in this simulation. However, the actual correlation between the item clusters was higher because all items share an overall random effect. Considering this common effect, the

correlations between the logits of the two clusters are .5 and .83 for a covariate weight correlation of 0 and .5, respectively.

In order to investigate the effect of the associations induced by the intercept, additional data sets were generated with a model as in Equation 9, but without a random intercept (without θ_p). This formulation of the model leads to a special case of the multidimensional two-parameter logistic model (2PL), with the discrimination parameters being restricted to the values of 0 and 1 and without a general random effect. In the simulation, the same design factors were used as earlier.

The distribution of the adjusted Rand index values for the original data (including all sample sizes) and for the data without a common ability is shown in Figure 9. As expected, ADCLUS is clearly more effective in revealing the latent structure when the correlation between the item groups is smaller.

Insert Figure 9 about here.

The mean and the variance of the obtained adjusted Rand index value were 0.85 and 0.05, respectively. The ADCLUS procedure led to excellent recovery in 59.2% of the data sets, good recovery in 12% of the data sets, moderate recovery in 9.5% of the data sets and poor recovery in 19.3% of the data sets. When only the data sets with sample sizes from 500 on are considered, the mean of the adjusted Rand index values was 0.94 and the variance was 0.01. ADCLUS yields excellent, good, moderate and poor recovery in 78.6%, 10.8%, 5.6% and 5% of the data sets, respectively.

In sum, the presence of an overall ability somewhat deteriorates the performance of ADCLUS. Furthermore, it can be expected that the smaller the variance of the common random effect, the better the performance of ADCLUS, but this issue was not investigated here.

Determining the true number of hidden item properties

In practice, the number of properties is often unknown, and a decision rule is needed for deciding upon the number of hidden item properties to work with. The present section is devoted to this issue. The data with sample size of 100, 500 and 1000 from the original study are used for this purpose.

Given that ADCLUS can reveal the hidden item properties reasonably well, the proposed next step is to extract more item properties than the true number is, and to include all these item properties into the ALR model formulation in order to estimate the association weights. Two *ALR-based methods* are proposed. First, the significance of these estimates could be a basis for determining the number of hidden item properties. Second, an elbow criterion can also be used for the estimates of the property weights.

However, ADCLUS itself may be an effective method to determine the true number of hidden item properties, without any further modeling through ALR. In principle, the BIC measure can be used for deciding upon the number of clusters in ADCLUS, but in our case, there is not a similarity matrix available per person. As it has been already mentioned, according to Lee (2002) one can rely on expert opinion about estimating the data precision of the standardized similarity matrix expressed on a scale: precise, average, imprecise. The corresponding suggested precision numbers are .05, .10, .15, respectively. However, using these values, the correct number of item properties is rarely found for the data generated for

Study one. When larger precision numbers are considered, corresponding to less precise data, a precision number of .4 leads to reasonable results: the correct number of item properties is found for 27.5%, 59.2 % and 66.7 % of the data sets, with a sample size of 100, 500, and 1000, respectively. However, the appropriate prespecified precision number may vary in practice. Consequently, for the type of data under investigation, there is no obvious, straightforward method to define the pre value and to determine the number of properties.

Therefore, alternative methods are studied for ADCLUS. Because the ADCLUS weights should be very similar to the ALR weights, except for a scaling factor, they would not be a basis for a separate method. Two other *ADCLUS-based methods* are presented here. First, although there is not a similarity matrix available per person, the precision number might be derived from a random split of the data, so that two matrices of log-odds ratios are obtained, and a *pre* value can be determined. Second, the explained variance of different ADCLUS models with an increasing number of properties can be compared using an elbow criterion. A scree test is applied on the explained variance as the function of the number of item properties.

All in all, four methods were studied for revealing the true number of clusters: (1) the significance of the association estimates in ALR, (2) the elbow criterion for the ALR weights, (3) ADCLUS with a precision number derived from a random split, (4) the elbow criterion for the ADCLUS explained variance. Methods 1 and 2 are applied with and without cross-validation. The four proposed methods (not counting the methods with cross-validation as separate ones) are described in detail in the following.

Method 1 (ALR significance). Four item properties were extracted from ADCLUS for each data set. Using ALR with the four ADCLUS clusters, the significance values of the associations were used to decide which item properties do have an effect (α=.01). In each model, five association parameters were estimated, one for the intercept and one for each of

the four extracted item properties. Three of the association estimates are expected to be significant (two related to item covariates and one to the intercept). Overall, the procedure resulted in 78.6% hits and 21.4% missers. Unfortunately, this approach also yields 71% false alarms and only 29% correct rejections. The success rates for identifying the correct number of item properties are summarized in the first column of Table 1 (ALR significance). The number of item properties is mostly overestimated (for 66.9% of the data sets), and rarely underestimated (for 6.7% of the data sets), hence with a significance level of .05, the success rate is worse. In sum, the significance test largely overestimates the number of hidden item properties. Possible reasons for the overestimation are discussed later.

Insert Table 1 about here.

Method 2 (ALR weights). Based on the ordered association estimates from ALR, an elbow criterion was used to decide upon the number of latent item properties. For these analyses, six item properties were needed from ADCLUS to be comparable to the previous method, as will be explained. Following a principle explained by Ceulemans and Van Mechelen (2005), a *discrepancy difference measure* is calculated per weight; the discrepancy with the following weight is subtracted from the discrepancy with the previous weight. The first measure is therefore determined for the second weight and the last one is determined for the fifth weight. The maximal discrepancy difference is selected, as the elbow, and the number of properties is the order position of the selected weight minus one. As a result, the possible outcomes for six properties are one to four item properties, just as for the previous method.

The results are summarized in the column “ALR weights” in Table 1. As it can be concluded from the table, it is a more successful method than the previous one. However, even for the larger sample sizes (at least 500), the procedure gives correct decisions for only a bit more than half of the data sets. When all data sets are taken into account and the method led to incorrect decisions, the number of item properties was rather underestimated (30.3% of all data sets) than overestimated (22.2% of all data sets).

It is important to note that when the item properties are ordered based on their ALR weights, and when the sample size is sufficiently large (from 500 examinees on), the first two estimated item properties are mostly the ones most similar to the true item properties (measured by the adjusted Rand index). In case of a sample size of 500, and 1000, the estimated item properties which were the most similar to the true item properties come first in 75.8% and 84.2% of the data sets, respectively. The average adjusted Rand index values were 0.79 and 0.89 for these item properties compared to the true ones, respectively. This means that the ordering is mostly correct, but that the gap with the following weights is not sufficiently large to create an elbow effect.

Because the same data sets are used to extract the item properties by ADCLUS which are used to test these item properties with ALR, the Methods 1 and 2 capitalize upon chance. Therefore, a cross-validation was implemented. The data sets are randomly splitted and half of the data was used for cluster extraction with ADCLUS, and the other half for testing the weights of the extracted clusters with ALR. The results are shown in columns “Method 1 cross-validation”, and in “Method 2 cross-validation”.

Method 1 with cross-validation resulted in correct decisions about the number of the item properties for 10%, 46.7% and 45.8% of the data sets with a sample size of 100, 500, and 1000, respectively, when a significance level of .01 is used. The overall performance is better than it was without cross-validation, but not for data with a sample size of 100. Consequently,

although cross-validation may improve the performance in general, when the sample size is small, half of the data may not be enough for obtaining reasonable parameter estimates, and hence may weaken the results.

The performance of Method 2 is not improved by a cross-validation for these data. Actually, the percentage of correct decisions decreased for smaller sample sizes such as 100 and 500, when cross-validation was applied (see Table 1). However, for a sample size of 1000, the performance is slightly improved by cross-validation.

Method 3 (ADCLUS estimated pre). The precision value was also calculated from the data. All data sets were divided into two subsets and the log-odds ratio was calculated for each subset separately. From the obtained two similarity matrices the pre value could be calculated. However, based on the obtained pre values, the correct number of clusters could not be found for any of the data sets, as it is indicated in column "Estimated pre" in Table 1. The number of item properties was overestimated for all data sets.

Method 4 (ADCLUS variance accounted for). Following a principle explained by Ceulemans and Van Mechelen (2005) for fit statistics, *a discrepancy ratio* is determined for each property, as the ratio of the discrepancy with the previous versus the discrepancy with the following. The first ratio is the ratio of the discrepancy of the explained variance of one property with the explained variance of just an intercept (zero explained variance) versus the discrepancy of the explained variance of two properties with the explained variance of one property. The last ratio is the one with 4 minus 3 versus 5 minus 4. The elbow is identified by the maximum ratio, and so is the number of item properties. Because the ratio can be determined for the properties one to four, four outcomes are possible: one to four properties, as for the previous methods. In order to have four possible outcomes, five properties were needed.

The results can be seen in column 'VAF' in Table 1. The method performed well for large sample sizes, and it seems slightly better than the Method 2. The overall percentage correct is still not very high. Considering all sample sizes, when not the right number of item properties is chosen, the number is mostly underestimated (for 43.1% of the data), not overestimated (for 11.1% of the data).

As a basis for the evaluation of the results, the method of principal component analysis (PCA) was applied on the raw data, as an alternative for revealing the number of underlying dimensions. Although it is not an orthodox method for binary data, it has been successfully applied on similar data in a comparative simulation study (Balazs, Hidegkuti & De Boeck, 2006). For this simulation study, the eigenvalues were calculated up to six item properties and, in order to be consistent, the earlier described scree test, based on the discrepancy difference measure was applied to the eigenvalues. The PCA leads to a correct decision only in 26.7%, 27.5% and 33.3% of the data sets for a sample size of 100, 500 and 1000, respectively. Consequently, the ALR weights method (Method 2) and the ADCLUS explained variance (Method 4) are much better criteria for revealing the number of item properties than the PCA is.

The overall percentage of correct decisions (45.8%) is only moderate for the best methods (ADCLUS variance accounted for and ALR weights), but it varies a lot depending on the design factors. For example, considering Method 4, correct decisions were obtained for 24.2%, 48.3% and 65% of the data when the probability of the item property was .3, .5 and .7, respectively. The effect of the variance and the correlation was smaller. The recovery rate was 45% for a variance of 1 and 46.7% for a variance of 2 and it was 58.3% for uncorrelated random effects and 33.3% for correlated random effects. Considering data with a sample size of at least 500, the recovery rate is 73.3% for uncorrelated random effects and 39.2% for correlated random effects of the two properties. When the sample size is at least 500 and the

probability of the item properties is .7, the recovery rate is 82.5%. When data sets with a sample of at least 500 are considered, the effect of the variance is not prominent. However, when the sample size is 1000, the effect of the variance becomes remarkably large. The recovery rate is 30% for a variance of one, and 63.3% for a variance of two.

In sum, the results largely depend on the specific features of the data set. Besides, given the fact that many challenging data structures are included, the recovery rates are rather good, and certainly they are in comparison with PCA.

Note on the significance of the extracted property effects

Because the number of false alarms with the ALR significance method generated was remarkable, some additional analyses were carried out in order to understand this result. Apart from the approximate nature of ADCLUS to decompose log-odds ratios into additive sources of random effects, there are two possible reasons for this phenomenon (1) capitalization on chance, and (2) extraction of disjunctive clusters.

In order to test for the first possibility, a cross-validation was performed, as it has been reported earlier. Accordingly, although for Method 1 the cross-validation was useful from already a sample size of 500 on, for Method 2 it seems to be beneficial only from a sample of 1000 on. Implementing cross-validation improved the results for larger sample sizes, but Method 2 without cross-validation, and Method 4, still do perform better. In sum, capitalization upon chance can be a reason for the overestimation of the number of significant item properties when ALR is used after ADCLUS and the sample size is small.

The second possibility is that more than one item property is extracted related to the same true property, for example if the extracted properties are disjoint fragments of the same true property. If the algorithm produces disjoint clusters, it may be expected that a too high number of clusters is found. In order to test for the second possibility, stepwise generalized

Boolean regression analyses were used, as implemented in the Combination Rule Analysis program (CRA; Van Mechelen, 1990). The CRA program constructs logical combinations of binary variables to predict a binary criterion. The conjunctive and the disjunctive combinations were considered in the analyses. In 28.9% of the data sets, the prediction of a true property was more successful with a conjunctive combination of extracted properties than with any of the extracted item properties separately. In 17.1% of the data sets, the disjunctive combination of the extracted item properties was a more successful predictor of the true item properties than any of the extracted item properties separately. Because less disjunctive combinations were found, the second possibility (tendency to favor property fragments), must be rejected. However, because in 46% (28.9% +17.1%) a better prediction of the true properties was obtained, it seems that the additional extracted properties are related to the true ones nevertheless. This may be the consequence of the approximate nature of ADCLUS for log-odds as a way to detect sources of random effects, just as a PCA for binary data is an approximate method and yields a too high number.

Conclusions

Based on the analysis with two extracted properties for each data set, ADCLUS seems to work well for the recovery of the true properties, but a sample of 500 or higher is desirable for good results. The effectiveness of the method clearly depends on some important data features. The most challenging data sets are those with a few items per property, with a small variance of the random effects combined with a high correlation between the random effects. When the sample size is 1000, when the probability of the item properties is 0.7, when the variance is 2, and when the correlation is zero, the mean adjusted Rand index is 0.97.

In order to investigate the possibility discussed by Bechger, Verstralen, and Verhelst (2002) for the LLTM, that different solutions may exist, in this case for ADCLUS, both the set of two true properties, and the first two extracted properties are regressed on the log-odds ratios. When the item clusters were not revealed perfectly, that is, when the Rand index was smaller than one, the two multiple correlations were never nearly identical, hence the possibility of equivalent solutions must be rejected.

Determining the correct number of properties seems a difficult challenge. The proposed method of testing the significance of the property weights using ALR generated a large number of false alarms. The reason is partly capitalizing on chance. The extracted item properties result in an overestimation of the number of properties with a significant weight, so that the number of properties is overestimated in comparison with the true number. When the sample size is large (from 500 on), cross-validation can improve the recovery. Partly, the reason of the overestimation of the number of the item properties when using ADCLUS is that the dependency of the association structure on the structure of the means is not taken into account, as explained earlier.

However, the best investigated methods seem to be the discrepancy difference method applied on the ALR weights and the discrepancy ratio method applied on the variance accounted for using ADCLUS. The performance of the ALR weights may be further improved by implementing cross-validation when the sample size is large (at least 1000). The results vary largely as a function of the data features (design factors). They are clearly better for a combination of a large sample size and uncorrelated random effects of the properties.

Finally, using a derived precision value for ADCLUS does not seem to work at all, whereas it is the recommended method from the point of view of the algorithm.

Study two

In a second simulation study, the performance of ADCLUS was compared to DETECT for revealing the underlying data structures with non-overlapping and overlapping item properties, and also a combination of the two methods was studied.

DETECT (Dimensionality evaluation to enumerate contributing traits; Kim, 1994; Stout, Habing, Douglas, Kim, Roussos & Zhang, 1996) is a method for revealing data structures with hidden non-overlapping dimensions. It is a nonparametric method, which is based on the conditional covariance of the item pairs (conditional on the general dimension).

In the case of a simple structure, DETECT is expected to provide the correct partitioning of the items. 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). For an approximate simple structure with clearly separable item subgroups, the method is expected to give a good estimation of the partitioning. The DETECT procedure provides the R index for identifying simple structure, a value greater than .8 indicating approximate simple structure, a value of one simple structure. An important condition for the validity of the DETECT index is that the structure is at least an approximate simple structure; therefore when the R value is smaller or equal to .8, the DETECT results should not be taken into account.

Using DETECT, a structure is considered unidimensional if the DETECT value is smaller than 0.1 according to the manual (The William Stout Institute for Measurement, 2003), or smaller than 0.2 according to some other authors (van Abswoude, van der Ark, & Sijtsma, 2004; Uribe-Zarain, Nandakumar, & Yu, 2005). However, for the data sets of Study two, the conclusions were not different for the two cut-off values. If the conclusion from DETECT is unidimensionality, then, in fact, the true number of properties is zero, because the one dimension is a general one and refers to the random intercept. If the conclusion from

DETECT is multidimensionality, DETECT results in a partition with at least two clusters, and hence, the corresponding number of properties must be at least two. Consequently, there is no room for one property in addition to a general dimension, since this kind of structure would certainly not be a simple structure. From two on, the number of clusters should directly reflect the number of properties. Only the maximal number of clusters can be specified by the user, and the algorithm determines the optimal number within the range between two and the specified maximum.

Of course, it can also be expected that DETECT does not perform so well for overlapping cluster structures as for a partition, because overlapping clusters contradict simple structure. Nevertheless, it is interesting to compare the performance of the two methods for both kinds of structures, since in practice it is not known in advance whether the true clusters overlap. In theory, when the item clusters do not overlap, both methods may be suitable; but when the item clusters do overlap, ADCLUS should be more appropriate, although DETECT can still be considered as an approximate method. Additionally, the performance of a combined method is investigated. In the combined method, the R value from DETECT is used to indicate the presence of absence of simple structure. When simple structure is indicated, the DETECT results are accepted, otherwise ADCLUS is applied on the data, and the ADCLUS results are considered.

The model with two covariates in Equation 9 was used for data generation, but this time the covariates were either overlapping or non-overlapping, and their value was not determined in a stochastic way. When the structure was *non-overlapping*, the first half of the 24 items had item property one and the second half had item property two; or the first 8 items had item property one and the remaining 16 items had item property two. When the covariates had an *overlapping* structure, the first eight items had item property one, the next eight items

had both item properties, and the last set of eight items had item property two. Hence, three structures were used, two non-overlapping ones and an overlapping one.

The effects of both covariates had a random distribution. The intercept and both random weights were normally distributed with a zero mean. The variance of the intercept was one. The variance of the covariate weights was either 1 or 2. The correlation of the three random effects was either zero or .5. The overall mean success probability was .5. The sample size was 100, 500 or 1000. The three basic structures, the two variance values for the random weights of the covariates, the two correlation values among the random effects, and the three sample sizes were the factors of a fully crossed design. Again, ten data sets were generated per design cell, so that in total 360 data sets were generated.

Both, ADCLUS and DETECT were applied to these data. Another possible approach is using ADCLUS in combination with DETECT which was also studied. If the R index is not higher than .8, the ADCLUS results should be chosen. If the R value is higher than .8, the DETECT result should be chosen. For all three procedures (ADCLUS, DETECT and the combined method) the extracted clusters (given the true number of clusters) were compared to the true ones by calculating the adjusted Rand index and it was also investigated whether the correct number of item properties is found.

Results

Also for Study two it is investigated to what extent the true item properties can be recovered when the true number of item properties are extracted. Please note that DETECT may not provide results for all data sets, since the lack of simple structure implies that the results should not be considered.

First, the results are discussed for *non-overlapping* properties with an equal number of items per property. For data sets *with equally distributed items (12/12)*, DETECT indicated at least an approximate simple structure for 30%, 87.5% and 100% of the data sets for a sample size of 100, 500, and 1000, respectively.

In case of non-overlapping properties, two item clusters correspond to the item property patterns (01, 10). Therefore, DETECT was used to obtain two item clusters, and the adjusted Rand index was derived from a two-by-two contingency table. But, because ADCLUS permits the clusters to overlap, two clusters can yield four property patterns (11, 10, 01, 00), and hence, a four-by-two contingency table is used for the recovery evaluation of ADCLUS.

The extracted item properties had to be linked to the true item properties. For that reason, the similarity of each extracted item property to each true item property was expressed by the adjusted Rand index. The sum of the adjusted Rand index values corresponding to all possible combinations of the extracted and true item cluster pairs were calculated. The combination with the highest sum of adjusted Rand indices was selected.

The mean of the adjusted Rand index values for the three methods (DETECT, ADCLUS and combination of the two) are shown in Table 2 for each sample size. For DETECT, the mean Rand index value is only calculated for those data sets where approximate simple structure was indicated. Based on these results, DETECT was more successful in revealing the true structure than ADCLUS, however the difference is not large when the sample size is 500, and not remarkable when the sample size is 1000. Furthermore, one should note, that for 41.3% of the data with a sample size smaller than 1000, the DETECT procedure indicated lack of simple structure, while it did not occur for a sample size of 1000. The combined method resulted in an improved performance in comparison with

ADCLUS. Considering that the combined method provides results for all data sets, this seems the best performing method.

Insert Table 2 about here.

The second issue is whether the procedures find the true the number of latent properties. For deciding on the number of item properties, DETECT was implemented with a maximal number of 12 clusters. Therefore the possible conclusions are zero and from two to twelve clusters. The solution zero is equivalent with a one-dimensional structure. The procedure never resulted in a number of clusters larger than three. Using ADCLUS with a maximum number of five clusters, the number of properties was determined based on the discrepancy ratio method applied to the variance accounted for (as for Method 4 in Study one). The possible conclusions from ADCLUS are from one to four item clusters.

As it can be seen in Table 3, the combined method was the most effective for revealing the true number of latent item properties. The performance of ADCLUS seems to be the second best. However, one should note that, when DETECT did not find the correct number of item properties, it either gave the wrong conclusion or no conclusion could not be drawn based on DETECT because the R value was not sufficiently large.

Insert Table 3 about here.

It was also investigated whether the inequality of the dimensions plays a role. Therefore, the three methods were also compared for non-overlapping data structures with

unequally distributed items (16/8). Again, the item properties define either a 01 or 10 pattern, hence the adjusted Rand index values and the estimations of the number of item properties are obtained in the same way as before for the determination of the similarity of the extracted properties to the true ones.

DETECT indicated at least approximate simple structure for 22.5%, 77.5% and 92.5% of the data sets for a sample size of 100, 500, and 1000, respectively.

Again, DETECT extracted clusters most similar to the true ones (see Table 2), but still for many data sets DETECT did not provide estimates. From the other two methods, the combined method performed slightly better than ADCLUS.

Both DETECT and ADCLUS were applied in the same way for determining the true number of item clusters, as for equally distributed items. As can be seen in Table 3, the combined method performed as the best. From the other two methods, DETECT performed somewhat better than ADCLUS in terms of percentage of correct decision on the number of item clusters. Again, it should be noted that if DETECT does not find the number of properties, it is either because the conclusion on the number of properties is wrong or because no conclusion could be drawn because of lack of simple structure.

The performance of all methods was worse for unequally distributed item than for equally distributed items. From a closer inspection of the data, it can be stated, that the performance of ADCLUS became worse especially for correlated data.

It has been also considered that the comparison between ADCLUS and DETECT, as far as based on the Rand index, may not be fair in the way as it was made, because the contingency tables differ in their size: four-by-two for ADCLUS, and two-by-two for DETECT. In order to make a fair comparison, the adjusted Rand index values were recalculated for the first and for the second covariate separately. In this way, two two-by-two

contingency tables were available both for DETECT and ADCLUS. The results were slightly better for ADCLUS than before, but the basic conclusions remain the same.

Finally, the performance of the three methods for *overlapping clusters* were also investigated. DETECT correctly indicated lack of simple structure for 70% of the data with overlapping structure.

The true structure is one with three item subgroups, corresponding to the patterns 11, 10 and 01. Therefore, when the agreement of the extracted and true item clusters was investigated, DETECT was used with a prespecified number of three clusters, so that DETECT would not be handicapped for the recovery of the true structure with three patterns. This results in a three-by-three contingency table to determine the adjusted Rand index. On the other hand, because two overlapping clusters estimated with ADCLUS can yield four patterns (11, 10, 01, 00), corresponding to the true properties, a four-by-three contingency table was used to determine the adjusted Rand index for ADCLUS.

The adjusted Rand index values are the highest for ADCLUS (see Table 2), but the combined method performed similarly well. Even in those cases when DETECT indicated simple structure, the adjusted Rand index values are much worse for DETECT than for the other two methods. It is not surprising, given the fact that DETECT is not a valid method, since its aim is to detect non-overlapping clusters. In general, in each design cell for DETECT, the means of the adjusted Rand index values are mostly about half of the values for non-overlapping covariates.

In order to obtain results from contingency tables with an equal size for both methods, the adjusted Rand index values were recalculated, separately for the two covariates, and hence, with two-by-two contingency tables. Again, the obtained results were not remarkably different from the original results.

For the determination of the true number of properties, the methods were used in the same way as before: DETECT was allowed to extract maximum 12 clusters, but it never extracted more item clusters than three; and ADCLUS was allowed to chose for maximum four item clusters (five item properties were extracted and the discrepancy ratio method was used). Based on the idea behind DETECT, the items with the 00 and 11 patterns may appear randomly in clusters based on the 01 and 10 patterns, or the 11 pattern may appear in a separate cluster.

ADCLUS performed better than DETECT and the combined method in terms of determining the correct number of item properties for data with overlapping properties, as it can be seen in Table 3. The combined method performed worse for sample sizes of 500 and 1000. The performance of DETECT was even worse. Although DETECT is not developed for this kind of structure, when DETECT is used in practice, the user would not know whether the properties overlap. Therefore, it is still meaningful to investigate how many clusters DETECT derives when the adjusted Rand index value indicates approximate simple structure.

Conclusions

In general, for the investigated three data structures the combined method seems to perform the best. The performance of ADCLUS is somewhat worse for data with non-overlapping item clusters, but it is somewhat better for data with overlapping clusters.

When the item properties are non-overlapping, DETECT performs well, but it often does not provide conclusions about the data structure, when the sample size is small and its performance is poor for data with overlapping item clusters. In practice, item properties can be overlapping, and in such cases ADCLUS and the combined method perform better. When the item covariates are non-overlapping, it is usually the result of a carefully and successfully

controlled item structure with item properties which are known in advance to the analyses, and with a categorical design. Note that an orthogonal design does yield overlapping properties.

Therefore, it may be recommended to use the combination of DETECT and ADCLUS whenever it seems reasonable that the underlying item properties may be non-overlapping, whereas ADCLUS is to be recommended when one knows that the properties overlap.

General conclusions

Based on the results, ADCLUS can be effectively used for the detection of unknown underlying binary item properties which are sources of individual differences in binary, person by item data. It was demonstrated in Study one, that given the true number of properties, ADCLUS may find the correct underlying structure, but for a good performance, a sample size of at least 500 is needed. For deciding upon the number of item clusters, the discrepancy ratio method applied on the variance accounted for using ADCLUS and the discrepancy difference method applied on the ALR weights performed the best. The performance may be further improved by implementing cross-validation when the sample size is large, but for small sample sizes half of the data is not enough for a good performance of the method, hence cross-validation actually deteriorates the performance. The effectiveness of the methods varies depending on the design factors. However, the ADCLUS based method must at best be considered an approximate method.

In Study two, the performance of three methods: ADCLUS, DETECT and a combination of the two was compared for three kinds of data structure. Based on Study two, and considering data with non-overlapping item clusters, a combination of ADCLUS with DETECT seems to perform better than ADCLUS or DETECT separately. However, when

overlapping item-clusters are expected ADCLUS performs better. When the kind of structure, overlapping or not, is not known, the combined method is a good choice.

References

Arabie, P. & Carroll, J. D. (1980) MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 2,* 211-235.

Balázs, K., Hidegkuti, I., & De Boeck, P. (2006). Detecting heterogeneity in logistic regression models. *Applied Psychological Measurement,30,* 322-344.

Bechger, T.M., Verstralen, H.H.F.M., & Verhelst, N.D. (2002). Equivalent linear logistic test models.*Psychometrika, 67*, 123–136.

Barnhart, H. X. & Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses, *Biometrics, 54*, 326-335.

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

Blumer, H. (1969). Symbolic Interactionism: Perspective and Method. Englewood Cliffs, NJ: Prentice-Hall.

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

Carroll, J. D. & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika, 2*, 157-169.

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

Chaturvedi, A., & Carroll, J. D. (1994). An alternating combinatorial optimalization approach to fitting the INDCLUS and generalized INDCLUS models, *Journal of classification, 11,* 155-170.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/ psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S.

Egan, D. E. (1979). Testing based on understanding: Implications from studies of spatial ability. *Intelligence*, 3, 1-15.

Embretson, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Embretson, S. E. (1985). Test Design: Developments in Psychology and Psychometrics. London: Academic Press.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research, *Acta Psychologica, 37,* 359-374.

Fowlkes, E. B. & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78,* 553-569.

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

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champain.

Hojo, H. (1983), A maximum likelihood method for additive clustering and its applications, *Japanese Psychological Research, 25*, 1991-201.

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

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. van Duin & T. A. B. Snijders (Eds.), *Essays on item response theory.* New York, NY:Springer-Verlag.

Junker, B. W., & Sijtsma, K. (2001) Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-273.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773-795.

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

Lawson, C. L. & Hanson, R. J. (1974). Solving *least squares problems.* Englewood Cliffs, NJ:Prentice-Hall.

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.

Lee, M.D. (2002). A simple method for generating additive clustering models with limited complexity. *Machine Learning, 49, 39-58.*

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.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523-547.

McKinley, R. J.(1989). *Confirmatory analysis of test structure using multidimensional item response theory* (Research reports No. RR-89-31). Princeton, NJ: Educational Testing Service.

Mirkin, B. G. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices, *Journal of classification, 4,* 7-31.

Mirkin, B. G. (1989). Erratum, *Journal of Classification, 6,* 271-272.

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

Pan, W. (2001). Model selection in estimating equations. *Biometrics. 57*, 529-534.

Pervin, L.A. (1977). The representative design of person-situation research. In D. Magnusson & N.S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*.

Prentice, R. L. & Zhao, L. P. (1991). Estimating equations for parameters in the mean and covariates of multivariate discrete and multivariate responses. *Biometrics. 47*, 1033 1048.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

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

Segall, D. O. (2001). General ability measurement: an application of multidimensional item response theory. *Psychometrika*, 66, 79-97.

Shepard, R. N. & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 2,* 87-123.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods, 3,* 386-396.

Sternberg, R.J. (1977a). Intelligence, Information Processing, and Analogical Reasoning. Hillsdale, NJ: Erlbaum.

Sternberg, R.J. (1977b). Component processes in analogical reasoning. *Psychological Review, 34,* 356-378.

Sternberg, R. J. (1979). The nature of mental abilities. *American Psychologist, 34*, 214-230.

Stout, W. F., 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.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. L. Glaser, A. M. Lesgold & M. G. Shafto (Eds.), *Diagnostic Monitoring*

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.

Tenenbaum, J. B. (1996). "Learning the structure of similarity", in Advances in Neural Information Processing Systems, Volume 8., downloaded from: http://web.mit.edu/cocosci/Papers/nips95.pdf

Van Mechelen (1990). A FORTRAN program for the detection of logical relations between a set of predictors and a criterion variable. *Multivariate Behavioral Research, 25,* 207 209.

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.

Whitely, S. E. (1978). Information-processing on intelligence test items. Some response components. *Applied Psychological Measurement, 1,* 465-476.

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

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

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

Zhao, L. P. & Prentice, R. L. (1990). Correlated binary egression using a quadratic exponential model. *Biometrica. 77*, 642-648.

Zheng, B. (2000). Summarizing the goodness of fit of generalized linear models for longitudinal data. *Statistics of medicine, 19,* 1265-1275.

Ziegler, A., Kastner, C., & Blettner, M. (1998). The generalized estimation equations: an annotated bibliography. *Biometric Journal. 40*, 247-262.

Table 1

Percentage of correct decisions on the number of latent properties

Sample size	Method 1 ALR Significance $\alpha<0.01$	Method 2 ALR Weights	Method 1 Cross-validation	Method 2 Cross-validation	Method 3 ADCLUS Estimated pre	Method 4 ADCLUS VAF
100	28.3	33.3	10.0	29.2	0	25.0
500	25	50.8	46.7	36.7	0	50.8
1000	25.8	58.3	45.8	60.0	0	61.7

Table 2

Average adjusted Rand index values

Structure	Non-overlapping, equal (12/12)			Non-overlapping, unequal (16/8)			Overlapping (8/8/8)		
Sample size	100	500	1000	100	500	1000	100	500	1000
DETECT	0.99	1	1	0.96	0.99	1	-	0.48	0.56
ADCLUS	0.57	0.89	0.96	0.50	0.84	0.95	0.44	0.88	0.96
Combined	0.60	0.93	1	0.50	0.89	1	0.44	0.73	0.71

Table 3

Percentage of correct decisions on the number of properties

Structure	Non-overlapping, equal (12/12)			Non-overlapping, unequal (16/8)			Overlapping (8/8/8)		
Sample size	100	500	1000	100	500	1000	100	500	1000
DETECT	27.5	87.5	100	20.0	77.5	92.5	0.0	10.0	25.0
ADCLUS	60.0	85.0	100	22.5	65.0	82.5	45.0	82.5	90.0
Combined	67.5	92.5	100	32.5	82.5	92.5	45.0	65.0	55.0

Figure captions

Figure 1

The adjusted Rand index values as a function of the sample size

Figure 2

The adjusted Rand index values as a function of the probability of the covariates

Figure 3

The adjusted Rand index values as a function of the variance of the random effects

Figure 4

The adjusted Rand index values as a function of the correlation of the random effects

Figure 5

The adjusted Rand index values as a function of the correlation of the random effects and the sample size

Figure 6

The adjusted Rand index values as a function of the variance of the random effects and the sample size

Figure 7

The adjusted Rand index values as a function of the probability of the item properties and the variance of the random effects

Figure 8

The adjusted Rand index values as a function of the correlation of the random effects and the variance of the random effects

Figure 9

The distribution of the adjusted Rand index values for data sets with a general random effect (left panel) compared to data sets without a general random effect (right panel)

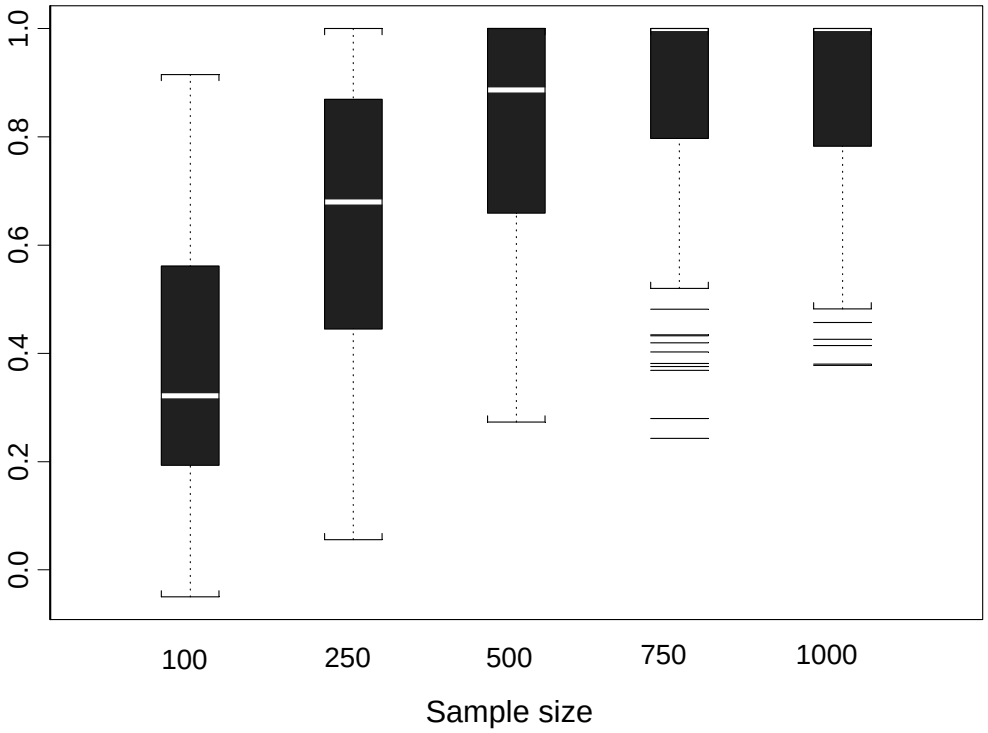


Figure 1

The adjusted Rand index values as a function of the sample size

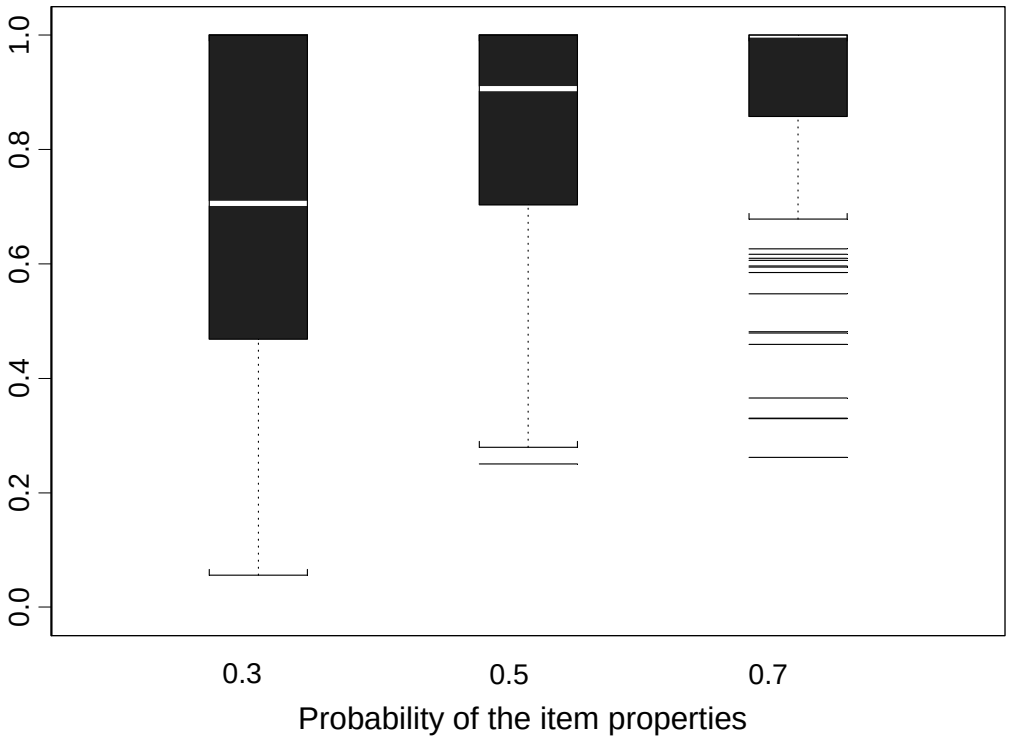


Figure 2

The adjusted Rand index values as a function of the probability of the covariates

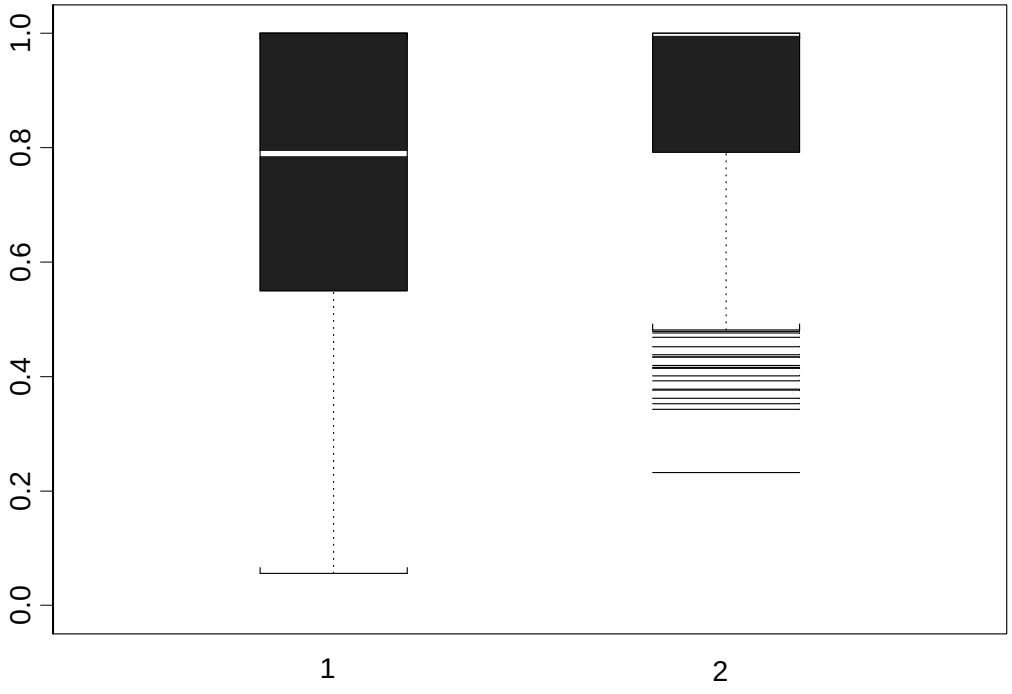


Figure 3

The adjusted Rand index values as a function of the variance of the random effects

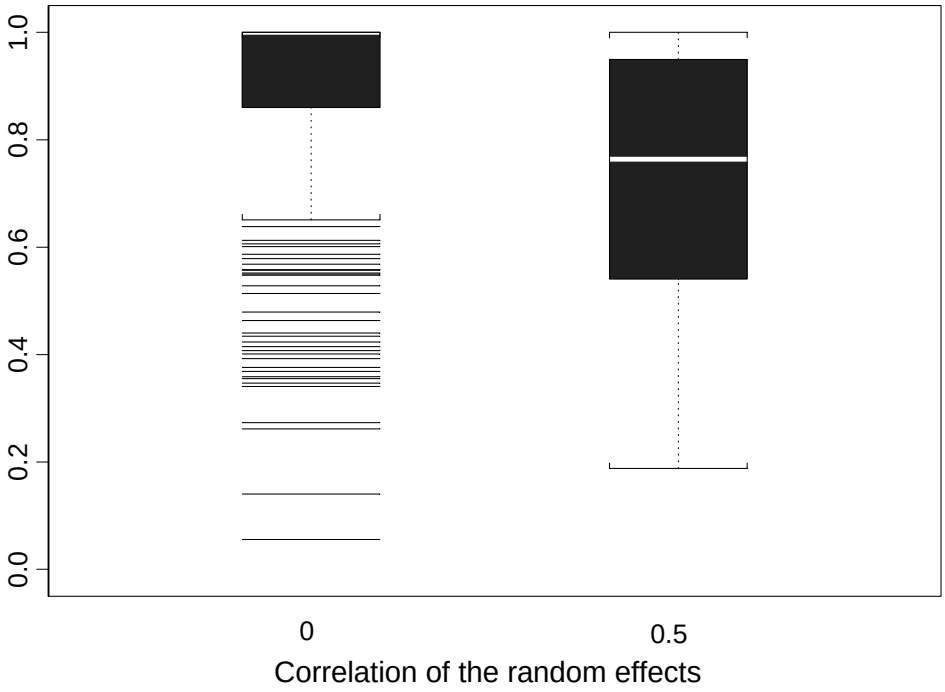


Figure 4

The adjusted Rand index values as a function of the correlation of the random effects

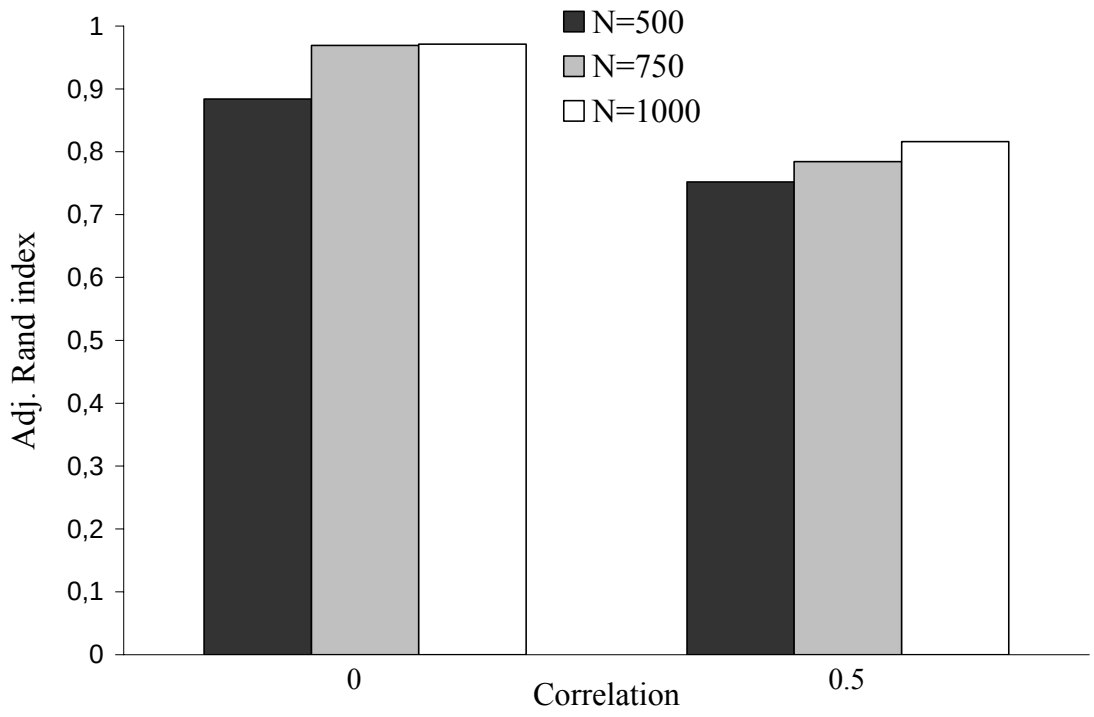


Figure 5

The adjusted Rand index values as a function of the correlation of the random effects and the sample size

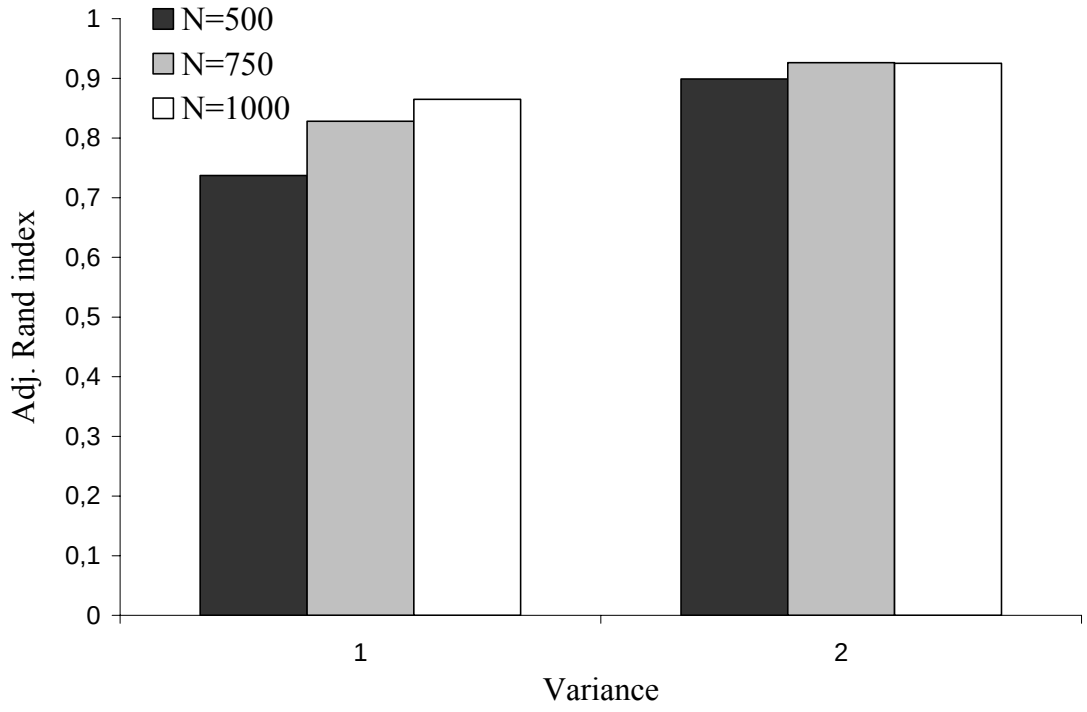


Figure 6

The adjusted Rand index values as a function of the variance of the random effects and the sample size

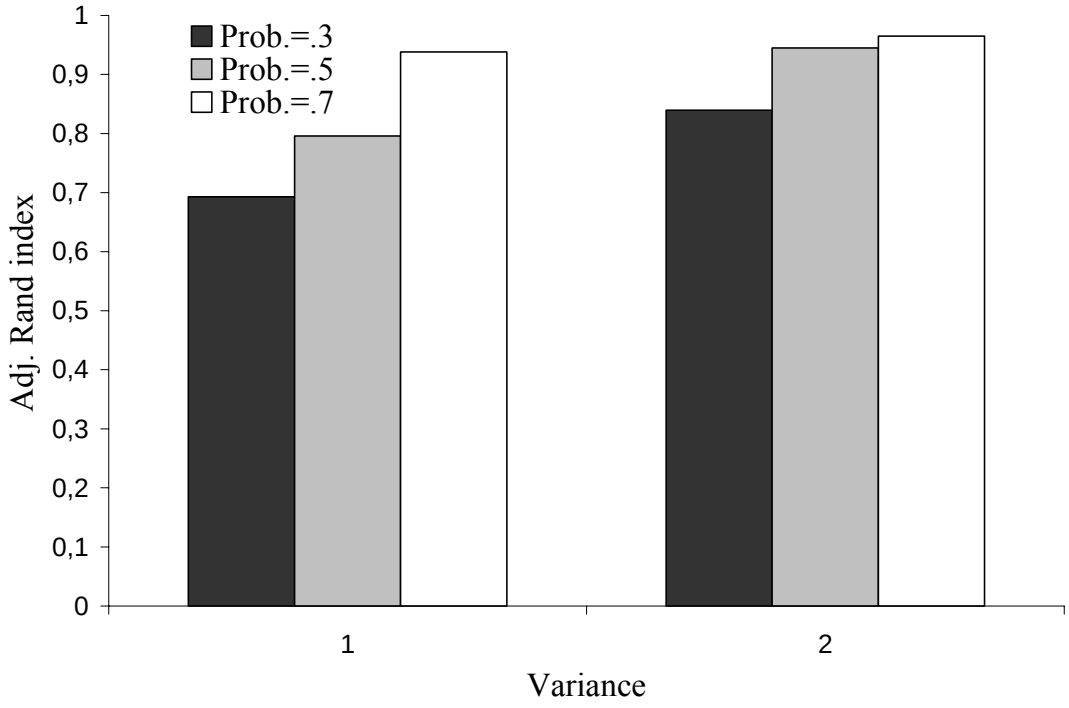


Figure 7

The adjusted Rand index values as a function of the probability of the item properties and the variance of the random effects

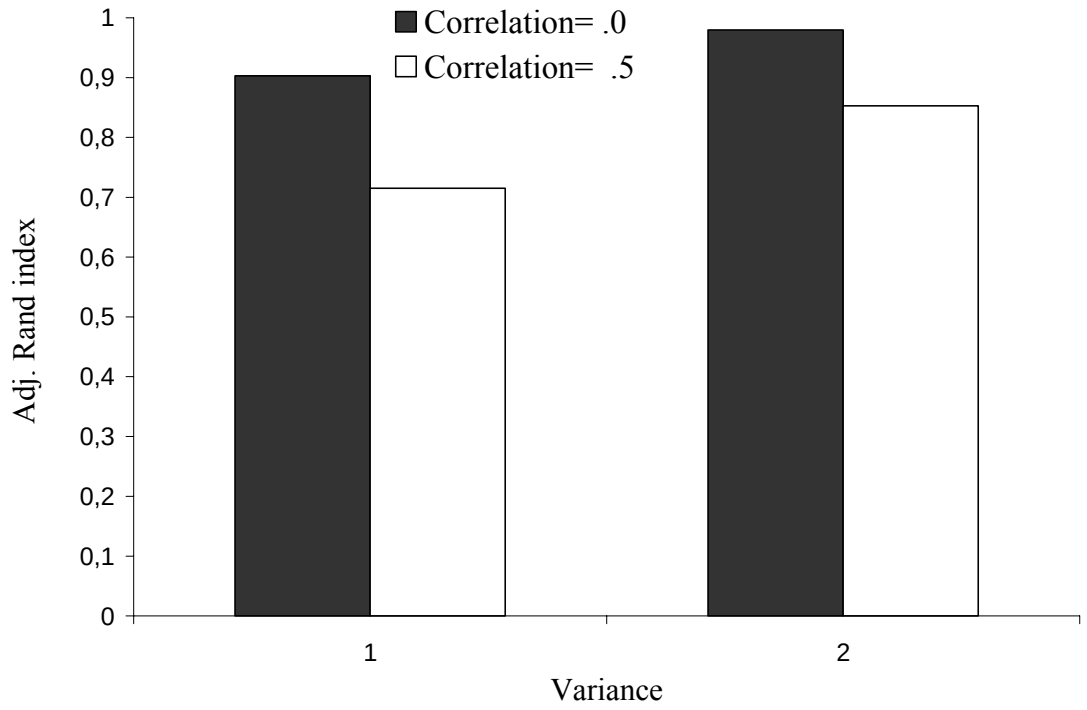


Figure 8

The adjusted Rand index values as a function of the correlation of the random effects and the variance of the random effects

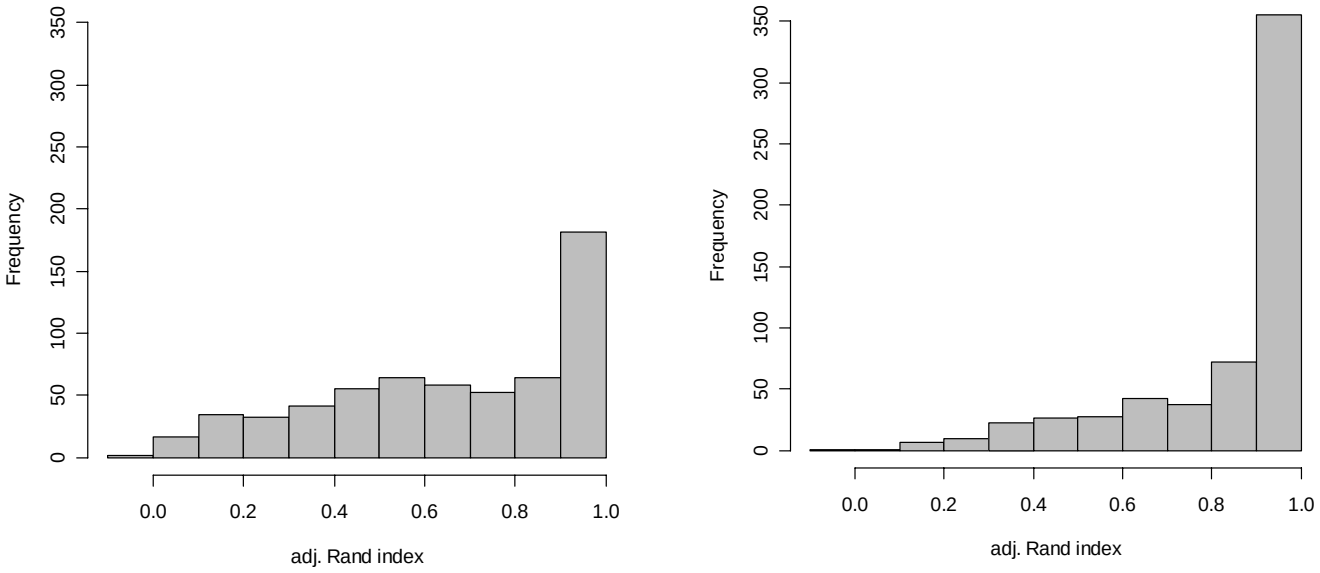


Figure 9

The distribution of the adjusted Rand index values for data sets with a general random effect (left panel) compared to data sets without a general random effect (right panel)