Behaviormetrika
Vol.35, No.1, 2008, 35–54

CLUSTERING NOMINAL DATA WITH EQUIVALENT CATEGORIES

Marian Hickendorff*, Willem J. Heiser*, Cornelis M. van Putten*, and
Norman D. Verhelst**

The problem considered in the present paper is how to cluster data of nominal measurement level, where the categories of the variables are equivalent (the variables are replications of each other). One suitable technique to obtain such a clustering is latent class analysis (LCA) with equality restrictions on the conditional probabilities. As an alternative, a less well known technique is introduced: GROUPALS. This is an algorithm for the simultaneous scaling (by multiple correspondence analysis) and clustering of categorical variables. Equality restrictions on the category quantifications were incorporated in the algorithm, to account for equivalent categories. In two simulation studies, the clustering performance was assessed by measuring the recovery of true cluster membership of the individuals. The effect of several systematically varied data features was studied. Restricted LCA obtained good to excellent cluster recovery results. Restricted GROUPALS approximated this optimal performance reasonably well, except when underlying classes were very different in size.

1. Introduction

To identify a possibly existing group structure in a data set, we need clustering techniques. Clusters are formed such that objects or individuals in the same group are similar in some respect and different from objects or individuals in other groups (e.g., Arabie & Hubert, 1996). Aim of the current study is to discuss techniques that can obtain a clustering for one specific type of data, characterized by two features. Firstly, the data are categorical: all variables are of a nominal measurement level. Secondly, all variables have "equivalent" categories: the interpretation of the categories is identical and the variables can be considered replications of each other.

Such a problem can be encountered in many practical instances. For example, in research on mathematics education, Van Putten, Van den Brom-Snijders, and Beishuizen (2005) studied the strategies students used to solve a set of division problems. There are reasons to expect clusters of students characterized by their own specific strategy use. All variables (division problems) are categorical, with categories coding the different strategies students applied. Furthermore, applying a certain strategy has the same meaning for all problems. So, under a further assumption that task characteristics are not affecting strategy use, the variables can be considered replications. We need a suitable clustering

Key Words and Phrases: classification, latent class analysis, multiple correspondence analysis, GROUPALS, categorical variables, equality restrictions, equality constraints, parallel items

∗ Leiden University Institute for Psychological Research, The Netherlands
∗∗ CITO, National Institute for Educational Measurement, Arnhem, The Netherlands
Mail Address: E-mail: hickendorff@fsw.leidenuniv.nl

The research was supported by a grant for Master's Students in Psychometrics from CITO, National Institute for Educational Measurement. Requests for reprint may be sent to Marian Hickendorff, Division of Methodology and Psychometrics, Leiden University Institute for Psychological Research, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

technique to identify these subgroups of students.

Another example is in research on cognitive development. Often developmental theories define several qualitatively different stages in development characterized by a specific cognitive behavior. Tests are used to measure the individual's stage of development at a certain point in time, and consist of several tasks scored correct or incorrect. In some domains, the tasks can be assumed to be equally difficult, such as in research on analogical reasoning (Hosenfield, Van der Maas, & Van den Boom, 1997) or in research on the water-level task (Thomas & Hettmansperger, 2001). So, the data are categorical (specifically, they are dichotomous) and the categories have the same interpretation in all variables (tasks): they are replications. A clustering technique is needed to find the several groups of children representing different cognitive stages.

Furthermore, in social psychology surveys or personality tests, (a set of) items may measure the same construct and be scored in the same format (e.g. *agree*, *disagree*, *don't know*). When subsets of respondents are expected with specific response tendencies, those can be identified by a clustering technique. Some final examples are the identification of clusters of individuals for which the same categorical variable is repeatedly measured over some time intervals, or in identification of clusters of objects that are ranked by several judges on some specific property (ranking data).

In the following, two techniques that are suitable for clustering categorical data are discussed: latent class analysis (LCA) and the perhaps less well known optimal scaling procedure GROUPALS. We elaborate how these techniques can deal with variables with equivalent categories by means of equality restrictions. In the next part, some simulation results are reported. Data were simulated firstly based on artificial model parameters, and secondly based on results from empirical analyses of data on strategies for solving mathematics items. It was assessed how several data conditions affect the clustering performance, and how the performance of restricted GROUPALS compares to the performance of restricted LCA.

2. Clustering categorical data

We distinguish two approaches to clustering: the criterion-based and the model-based techniques. In the criterion-based approach to clustering, typically a measure for the "goodness" of any proposed partitioning is defined, and the purpose of the clustering method is to find the partitioning that is optimal with respect to this measure. These methods are therefore also referred to as optimization techniques for clustering. One of the more popular of these procedures is the K-means algorithm (MacQueen, 1967), which iteratively relocates individuals between classes, until no further improvement of the optimization measure can be found.

In most of these clustering methods it is required to compute either similarities or Euclidean distances between the individuals (Arabie & Hubert, 1996). The measurement level of the variables is an important issue, because with nominal data the categories have no meaningful numerical values, making the derivation of similarities or distances not as straightforward as for numerical data. As a way to overcome that difficulty, it is possible to first use (multiple) correspondence analysis to derive numerical values for the categories (optimal scaling), and then use K-means on these derived spatial coordinates. Such a sequential analysis may be inappropriate, as has been noted by several authors (Chaturvedi, Green & Carroll, 2001; Vichi & Kiers, 2001), since (multiple) correspondence analysis as a data reduction technique may identify dimensions that do not necessarily contribute to the identification of the cluster structure of the data, or worse, may even obscure or mask this structure. However, attempts have been made to overcome this problem. Van Buuren and Heiser (1989) proposed a method called GROUPALS, in which scaling and clustering are done simultaneously, so that the solution is optimal to both criteria at the same time. A similar method has been proposed by Hwang, Dillon, and Takane (2006). In their method, it is possible to differentially weigh the clustering and scaling. Results from analyses of an empirical data set were very similar to the GROUPALS solution, however. Finally, Vichi and Kiers (2001) have proposed a similar simultaneous method for numerical data, so-called factorial K-means.

Apart from criterion-based approaches to cluster analysis, there are model-based or probabilistic clustering techniques (e.g. Arabie & Hubert, 1996). These techniques require that a statistical model is postulated for the population from which the sample is taken (Vermunt & Magidson, 2002). Specifically, it is assumed that the data are generated by a mixture of underlying probability functions in the population. When all variables are categorical, these models reduce to latent class analysis (Lazarsfeld & Henry, 1968; Goodman, 1974, 2002).

Summarizing, two techniques that can cluster categorical data are the model-based latent class analysis, and the optimal scaling method GROUPALS. In the next section, these methods will be studied in more detail and the equality restrictions that can account for equivalent categories will be discussed. The following notation is used: let H be the data matrix of the form N individuals by m categorical variables each with l_j $(j = 1 \dots m)$ categories, and let K denote the number of classes or clusters the individuals belong to.

3. LCA with equality restrictions

3.1 Basic concepts

The latent class model assumes an underlying latent categorical variable that can account for the covariation between the observed or manifest categorical variables (McCutcheon, 1987; Goodman, 2002). This assumption translates into the axiom of local independence, meaning that conditionally on the level of the latent class variable (named here T with K levels), the probability functions are statistically independent (McCutcheon, 1987). In the hypothetical case of two observed variables (A and B) and one latent variable T, local independence means that the latent class model can be expressed as follows:

$$\pi_{u\,v\,\kappa}^{A\,B\,T} = \pi_{\kappa}^{T} \cdot \pi_{u\,|\,\kappa}^{A\,|\,T} \cdot \pi_{v\,|\,\kappa}^{B\,|\,T}. \tag{1}$$

In words, $\pi_{u\,v\,\kappa}^{A\,B\,T}$ is the probability of an individual scoring category u on variable A, category v on variable B, and category κ on (latent class) variable T. This probability can be expressed as the product of the *latent class probability* π_{κ}^{T} and the *conditional probabilities* $\pi_{u\,|\,\kappa}^{A\,|\,T}$ and $\pi_{v\,|\,\kappa}^{B\,|\,T}$ of scoring category u on variable A and scoring category v on variable B respectively, both conditional upon belonging to latent class κ.

Although in LCA cluster membership of the individuals is not estimated directly, one can derive a partitioning by determining the posterior probability that an individual is in latent class κ, given its response pattern on the manifest variables and the estimated class and conditional probabilities. A conventional procedure is to assign an individual to the latent class for which it has the highest posterior probability (modal assignment).

3.2 LCA with equality restrictions

To adjust the LC model for data in which all the variables have equivalent categories, restrictions on the basic LC model can be imposed. Goodman (1974) already discussed several restricted latent structures. In the present study, the conditional probabilities are restricted to be equal over the variables. In the hypothetical example of two parallel categorical variables A and B, this equality restriction is as follows: $\pi_{q\,|\,\kappa}^{A\,|\,T} = \pi_{q\,|\,\kappa}^{B\,|\,T} = \pi_{q\,|\,\kappa}$, for all categories $q = 1, \ldots, l$ and for all classes $\kappa = 1, \ldots, K$. From now on, we let the term restricted LCA refer to the LC model with equality restrictions on the conditional probabilities.

A popular iterative algorithm to estimate parameters in LC models is the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). Mooijaart and Van der Heijden (1992) studied parameter estimation for LC models with equality restrictions, and derived the likelihood functions to be solved in the M-step. Vermunt (1997) implemented a uni-dimensional Newton algorithm to solve these equations in his program LEM, a general and versatile program for the analysis of categorical data.

Imposing restrictions on parameters in LC models has another advantage. In unrestricted LC models, the number of parameters to be estimated $(m \cdot K \cdot (l-1) + K - 1)$ increases very rapidly as the number of variables and/or the number of categories increases. For example, in fitting an unrestricted 5-class model on data with 10 variables, each with 7 categories, 304 parameters have to be estimated. Estimating this many parameters can result in identification problems, and in departure of the distribution of the likelihood statistics from a χ^2-distribution (Collins, Fidler, Wugalter, & Long, 1993). Restricting the conditional probabilities to be equal diminishes the number of parameters to only 34 $(K \cdot (l-1) + K - 1)$. So, imposing equality restrictions can be a resolution to obtain more stability and parsimony when the number of parameters gets too large.

4. GROUPALS with equality restrictions

4.1 Basic concepts

GROUPALS is a clustering method for variables of mixed measurement level, proposed by Van Buuren and Heiser (1989). The rationale of the technique is the simultaneous clustering of the individuals by a K-means procedure and scaling of the categories by an optimal scaling technique. In case of categorical variables, optimal scaling amounts to multiple correspondence analysis, also called homogeneity analysis or HOMALS (HOMogeneity Analysis by Alternating Least Squares).

In homogeneity analysis (Gifi, 1990), quantifications on p dimensions are derived for (a) individuals: matrix X of size N x p, and (b) categories: m Y_j-matrices of size l_j x p. These are obtained by minimizing the loss function

$$\sigma(X; Y_1, \ldots, Y_m) = \frac{1}{m} \sum_{j=1}^{m} tr(X - G_j Y_j)'(X - G_j Y_j) \tag{2}$$

over the object scores X and the m Y_j-matrices with category quantifications with an alternating least squares algorithm, usually with normalizations $X'X = \mathrm{I}$ and $1'X = 0$. The m G_j-matrices (N x l_j) are indicator matrices for each of the m categorical variables. An entry g_{iq} of an indicator matrix G is equal to one if individual i scores in category q, and zero otherwise.

In GROUPALS, an extra clustering restriction is inserted in loss function (2). All individuals in the same group should be at the same position (at the cluster mean) in the p-dimensional space: $X = G_c Y_c$, with G_c (N x K) the indicator matrix for cluster allocation of the individuals and Y_c the (K x p) matrix of cluster points[1)]. This addition results in the following GROUPALS loss function

$$\sigma(G_c; Y_c; Y_1, \ldots, Y_m) = \frac{1}{m} \sum_{j=1}^{m} tr(G_c Y_c - G_j Y_j)'(G_c Y_c - G_j Y_j) \tag{3}$$

which is optimized over G_c, Y_c and the m Y_j-matrices, also by an alternating least squares algorithm. The cluster allocation of the individuals, the positions of the clusters and the category quantifications result from a GROUPALS analysis.

4.2 GROUPALS with equality restrictions

If all variables in the data set have equivalent categories, the GROUPALS loss function (3) may be adjusted by requiring the category quantifications (of equivalent categories) of all variables to be equal. So, the category quantifications of, for example, category 1 of variable A are required to be equal to the category quantifications of category 1 of variables B and C, on all dimensions of the solution.

Thus, GROUPALS with equality restrictions on the category quantifications (from now on: restricted GROUPALS) is optimal scaling of individuals and categories, under the restriction that individuals in the same cluster are on the same position on all dimensions ($X = G_c Y_c$), and under the restriction that the category quantifications for all variables are equal ($Y_1 = \ldots = Y_m = Y$). The loss function of restricted GROUPALS then becomes

[1)] Note that the matrix with cluster points is called Y in the original discussion of GROUPALS by Van Buuren and Heiser (1989), but we call it Y_c. This was done to distinguish it clearly from the category quantifications matrices Y_j and Y.

$$\sigma(G_c\,;Y_c\,;Y) = \frac{1}{m}\sum_{j=1}^{m} tr(G_cY_c - G_jY)'(G_cY_c - G_jY). \tag{4}$$

The algorithm for minimizing loss function (4) consists of two steps: (a) estimating Y for fixed G_c and Y_c, and (b) estimating G_c and Y_c for fixed Y. These steps are alternated until the loss has reached a specified convergence criterion.

4.2.1 Estimating Y for fixed G_c and Y_c

Define $F = \sum_j G_j$, a frequency matrix of size N x l ($l_1 = \ldots = l_m = l$), with for all individuals the frequency of the categories $q = 1, \ldots, l$. Furthermore, define $D_j = G_j{}'\,G_j$ and $D = \sum_j D_j$. The m D_j-matrices are diagonal matrices (l x l) with the frequency with which each category of variable j is chosen summed over all N individuals, and D is a diagonal matrix (l x l) with the frequencies of the categories, summed over all N individuals and all m variables. By rewriting the loss function and setting the partial derivative with respect to Y equal to zero, the following equation to compute the optimal quantifications Y results:

$$\hat{Y} = D^{-1}F{}'G_cY_c. \tag{5}$$

So, the quantification of a category is the weighted sum of the scores from the individuals that chose that category (weighted by how many times the individual chose that category), divided by how many times that category was chosen by all individuals.

It worth noting that (5) is the same equation as the one used to compute the quantifications Y for fixed X in correspondence analysis by an alternating least squares algorithm (Heiser, 1981; Greenacre, 1984). Indeed, as noted by Gifi (1990) and by Van Buuren and De Leeuw (1992), minimizing the multiple correspondence analysis loss function with equality restrictions on the category quantifications (without cluster restrictions on the individual scores) is equivalent to carrying out correspondence analysis on the frequency matrix $F = \sum_j G_j$. So, the category quantification step in restricted GROUPALS is the same as the category quantification step in correspondence analysis on the frequency matrix F, if it is carried out by an alternating least squares algorithm. Furthermore, Greenacre (1988) developed an algorithm for clustering the rows (or columns) of a contingency table such as frequency matrix F. However, in contrast to our focus on partitioning methods, Greenacre (1988) constructed an agglomerative hierarchical method, based on merging rows (or columns) such that the Pearson χ^2-statistic of the collapsed table remains as high as possible.

4.2.2 Estimating G_c and Y_c for fixed Y

In line with Van Buuren and Heiser (1989), letting $Z = \frac{1}{m}\sum_j G_jY$ (unrestricted individual scores computed from the category quantifications Y), and inserting the identity $G_cY_c = Z - (Z - G_cY_c)$ into (4), the loss function can be split into additive components as follows:

$$\sigma(G_c; Y_c; Y) = \frac{1}{m}\sum_{j=1}^{m} tr(Z - G_jY)'(Z - G_jY) + tr(Z - G_cY_c)'(Z - G_cY_c). \tag{6}$$

To minimize this over G_c and Y_c, the first part is constant, so it is only the second part that has to be minimized, which is actually the problem of sum of squared distances (SSQD) clustering. The SSQD criterion can be minimized by the iterative K-means algorithm, resulting in the cluster allocation matrix G_c and the cluster means Y_c. Explicitly, (6) is minimal by setting $Y_c := ({G_c}' G_c)^{-1}{G_c}' Z$: The position of each cluster is the centroid of all individuals belonging to that cluster.

4.2.3 Transfer of normalization

In order to prevent the algorithm from making X and Y zero, either the individual scores X or the category quantifications Y need to be normalized. Due to two types of restrictions, computational inconveniences arise with the conventional normalizations $X'X = \mathrm{I}$ or $Y'DY = \mathrm{I}$. Van Buuren and Heiser (1989) therefore proposed a transfer of normalization procedure. The idea is to switch between both types of normalizations, while preserving the loss. We adopted the same procedure in the present restricted GROUPALS algorithm, and refer to Van Buuren and Heiser (1989) for a more extensive discussion of these normalization issues.

5. Comparing GROUPALS and LCA

5.1 Perspectives

The main difference between (restricted) GROUPALS and LCA is the postulation of an underlying model in LCA, which is not the case in GROUPALS. In LCA, it can be tested with likelihood statistics χ^2 and L^2 whether the data significantly depart form the model estimated. In addition, specific model features and the validity of restrictions that can be imposed on parameters can be tested statistically. Furthermore, in LCA the existence of information criteria (AIC and BIC) is very convenient for evaluating the fit, since it allows for comparison of models that are not nested, and parsimony of the model is also taken into account. Such statistical tests for the fit and specific model features are not possible in GROUPALS. The loss is computed, but no formal testing criteria for this loss are available, although they can be simulated with permutation procedures or other nonparametric statistics. Also, no information criteria are available in GROUPALS, where an increase in number of dimensions and/or number of clusters results in a decrease in loss, but in a less parsimonious solution. There are no formal criteria to choose the “best” solution. Another difference is that the clustering in GROUPALS is deterministic, or “crisp”, while in LCA it is probabilistic, or “fuzzy”, making it also possible to assess a degree of uncertainty of classification.

The interpretation of the solution also occurs on a different basis. In LCA, the estimated latent class and conditional probabilities can be used to characterize and interpret the classes. In GROUPALS, the dimensions of the solution can be interpreted by the category quantifications and next the clusters can be interpreted by the position of the cluster points on these dimensions. A graphical representation is possible for interpretational ease.

5.2 Limitations

As already mentioned, unrestricted LCA suffers from a proliferation of the number of parameters when the number of variables and/or the number of categories increases, possibly resulting in identification problems. A limitation specific for GROUPALS is that is has the tendency to produce spherical, equally sized clusters, inherent to the K-means algorithm (Van Buuren & Heiser, 1989; De Craen, Commandeur, Frank, & Heiser, 2006).

A common limitation of both techniques is the occurrence of local optima in the estimation algorithms. In GROUPALS, the incorporated K-means algorithm is well-known to suffer from local optima (e.g., Steinley, 2003). In LCA, the estimation algorithm can convergence to a local optimum of the log-likelihood function to be optimized. As yet, the best way to deal with local optima is to try several starting configurations and interpret the solution with the best fit. Another common disadvantage is that the user should specify the number of classes in advance (and in GROUPALS, the number of dimensions too) and no formal criteria exist to determine these. The best way to deal with this, except from theory regarding the data at stake, is to try several sensible numbers of clusters and compare the solutions.

6. Simulation Studies: Assessing Cluster Recovery

The theoretical discussion is now supplemented with simulation results. To assess clustering performance in several conditions, data sets with known cluster structure were generated with systematically varied data features. The recovery of true cluster membership of the individuals could be assessed as an indicator of clustering performance.

We generated data according to the latent class model, because we know of no other existing (more neutral) generation technique that results in clustered categorical data. Therefore, conditions were optimal for performance of restricted LCA. We first studied the effect of some features of the data that were expected to affect the degree of separation of the classes and hence also the clustering performance of restricted LCA. Specifically, the *number of classes* was expected to be negatively related to overall cluster recovery, while the *number of variables* and the *number of categories per variable* were expected to be positively related to cluster recovery, analogous to the findings of Chaturvedi et al. (2001) for unrestricted LCA. Furthermore, Chaturvedi et al. (2001) did not find an effect of *relative class size* on cluster recovery in LCA. Finally, *sample size* was varied to assess whether results deteriorated for smaller samples.

In addition, all generated data sets were also clustered by restricted GROUPALS. In this way we could compare the performance of restricted GROUPALS to the maximal possible performance of restricted LCA. Furthermore, we investigated whether effects of the data features were similar. There was no reason to expect different trends for the data features, except for the class sizes. Van Buuren and Heiser (1989) noted that the K-means clustering in GROUPALS tends to partition the data in clusters of roughly equal size. So, for restricted GROUPALS we expected lower cluster recovery with unbalanced class sizes than with classes balanced in size.

Two simulation studies were carried out, differing in the origin of model parameters: the conditional and class probabilities. In the first study, data sets were generated according to artificial model parameters to make results comparable across cells of the design. In the second study we based model parameters on results of latent class analyses of an empirical data set on solution strategies for solving mathematics problems, to make the situation more realistic.

6.1 Data Generation

The generation procedure consisted of two steps. For each individual, $m + 1$ random numbers between 0 and 1 were generated from a uniform distribution. In step 1, the first number determined what class the individual belonged to, by comparing the number to the cumulative class probability vector. In step 2, the remaining random numbers determined the respective categories scored on the variables one at a time, by comparing them to the cumulative conditional probability vector of the categories, conditional on membership of the class determined in step 1. This procedure was repeated for all individuals in a sample. Resulting data sets contained (a) the categories individuals scored on the variables, to be analyzed by the clustering techniques, and (b) the true cluster membership of those individuals, used to assess cluster recovery.

6.2 Cluster Recovery Index

It was possible to use an external criterion for the recovery of true cluster membership in this study, because there was information available on the cluster structure of the data, apart from the clustering process. Several indices for measuring the agreement between two partitions exist. In this study, we used the Hubert-Arabie adjusted Rand index (1985), which is the most desirable index for measuring cluster recovery (Steinley, 2004). This index corrects for chance agreement, and is therefore unaffected by the presence of unequal cluster sizes. It takes on a value of 0 when the partitions are chosen at random, and it is 1 when the partitions are identical. Steinley (2004) has proposed some guidelines for the interpretation of this adjusted Rand index. Values below .65 indicate poor recovery, values higher than .65 moderate recovery, higher than .80 good recovery, and values above .90 indicate excellent recovery.

6.3 Clustering Analyses

The algorithm for GROUPALS with equality restrictions was programmed in MATLAB. Since restricted GROUPALS is much influenced by local optima, which was also apparent in some test runs, all the restricted GROUPALS analyses were performed 200 times with different random starting partitions, and the solution with the best fit (the lowest loss) was chosen as the optimal solution.

The restricted LCAs were carried out with LEM (Vermunt, 1997). Since LCA also suffers from local minima, all the restricted LCAs were carried out 20 times with different starting values. This number turned out to be sufficient for avoiding local optima. The solution with the highest log-likelihood was retained as the optimal solution. From this solution, only the posterior probability for class membership for the individuals was saved, and individuals were assigned to the class for which they had the highest posterior probability.

In the analyses by restricted GROUPALS as well as in the analyses by restricted LCA, the number of clusters was set equal to the number of clusters in the population that was analyzed. In restricted GROUPALS, the number of dimensions was set at the maximum number of dimensions, which is the minimum of either $K - 1$ or $l - 1$ (Van Buuren & Heiser, 1989), so that as much information as possible from the data was retained.

6.4 Simulation Study I

6.4.1 Design

The following data features were varied systematically: the number of variables (5 and 10), the number of categories each variable has (5 and 7), the number of classes (3, 4 and 5), and the relative size of these classes (balanced and unbalanced). The balanced class size condition was operationalized as all classes being of equal size, while the unbalanced class size condition was operationalized as every class being two times as large as the class preceding it (so, the largest class always contained more than 50% of the individuals). The full crossing of these data features resulted in a 2 x 2 x 3 x 2 design with 24 different cells. For each cell of the design, 500 random samples ($N = 300$) were generated. Finally, to study the effect of sample size, some additional simulations were carried out with $N = 100$ and $N = 50$.

6.4.2 Setting the Model Parameters

The parameters of the LC population model needed to be specified. The class probabilities were part of the design of the simulation study: relative class size, balanced or unbalanced. The conditional probabilities were not part of the research design, so these had to be determined in some other way. To minimize the possibility that the amount of similarity between classes would confound with the effects of interest, care had to be taken to make the conditional probabilities comparable over all the cells of the design in a well-defined way. Therefore we decided to make all classes equally (dis)similar from each other. We defined the conditional probabilities in such a way that the Euclidean distance between the configurations of conditional probabilities of two classes was equal for all possible pairs of classes, with a value of $\sqrt{.45} = .6708$. For example, the squared distance between the configurations of class α and β was $d^2 = \sum_{q=1}^{l}(\pi_{q\,|\,\alpha} - \pi_{q\,|\,\beta})^2 = .45$, with categories $q = 1, \ldots, l$. So, one could say that the classes were lying on a simplex.

For each combination of number of categories and number of classes a configuration of conditional probabilities meeting the simplex-requirement had to be determined. This amounted to solving a set of quadratic equations for each of these cases separately. However, there were indeterminacies in the solution of these equation sets, so some conditional

Table 1: Conditional probabilities of Simulation Study I

	5 categories					7 categories						
class	1	2	3	4	5	1	2	3	4	5	6	7
α	.05	.10	.08	.30	.48	.04	.05	.07	.10	.12	.06	.56
β	.60	.05	.10	.08	.17	.52	.04	.05	.07	.10	.12	.10
γ	.20	.58	.05	.10	.08	.08	.54	.04	.05	.07	.10	.12
α	.05	.10	.08	.30	.48	.05	.05	.07	.10	.02	.56	.15
β	.60	.05	.10	.08	.17	.19	.05	.05	.07	.10	.02	.52
γ	.20	.58	.05	.10	.08	.23	.48	.05	.05	.07	.10	.02
δ	.13	.14	.55	.12	.05	.02	.10	.10	.55	.05	.07	.10
α	.05	.10	.10	.22	.53	.05	.05	.08	.10	.12	.03	.57
β	.59	.05	.10	.10	.16	.51	.05	.05	.08	.10	.12	.09
γ	.18	.57	.05	.10	.10	.06	.54	.05	.05	.08	.10	.12
δ	.12	.13	.55	.15	.05	.07	.10	.55	.05	.05	.08	.10
ϵ	.20	.14	.06	.60	.00	.05	.10	.13	.02	.57	.05	.08

probabilities were fixed to a specific value in advance. The derived conditional probabilities are displayed in Table 1.

6.4.3 Results

The complete simulation study consisted of 24 cells and for each cell 500 samples were generated, so a total of 24 x 500 = 12,000 samples or cases were generated and analyzed. Each case had a score on two (repeated measures) dependent variables: the cluster recovery by restricted LCA and the cluster recovery by restricted GROUPALS, both measured with the adjusted Rand index. These data were analyzed with a five-way (one within, four between) repeated measures ANOVA (Tabachnick & Fidell, 2001). Because the dependent variable was bounded by one, multivariate normality was likely to be violated. However, ANOVA is robust against that violation. Moreover, logarithmic and square-root transformations of the adjusted Rand index did not change the results.

Since the cells of the research design contained a large number of observations (500 samples per cell), effects were likely to reach significance quite easily: the design was overly powerful. To prevent interpreting significant but trivial effects of no practical importance, a measure is reported for the strength of association, to assess the relative importance of the effects tested for. In ANOVA this effect size measure is partial η^2, which Cohen (1988) characterized as *small* ($\eta^2 = .01$), *medium* ($\eta^2 = .06$) and *large* ($\eta^2 = .14$). To keep the discussion succinct, only effects of at least large size are reported.

Relevant means and standard deviations are displayed in Table 2. In Table 3, the results are reported for the effects of the between part of the design (effects of the data features on the cluster recovery averaged over the clustering techniques) and of the within part of the design (main effect of clustering technique, and interaction effects of clustering technique with the data features).

The mean overall cluster recovery was good with .80. As the first row of the within part of Table 3 shows, there was a significant and large effect (partial $\eta^2 = .67$) for the difference in cluster recovery between restricted GROUPALS (mean adjusted Rand index

Table 2: Means and standard deviations (in parentheses) of the adjusted Rand index in Simulation Study I

		restricted LCA	restricted GROUPALS	average
number of variables	5	.72 (.07)	.63 (.09)	.68 (.07)
	10	.94 (.03)	.89 (.05)	.91 (.04)
number of categories	5	.82 (.12)	.74 (.15)	.78 (.13)
	7	.84 (.12)	.79 (.14)	.82 (.13)
number of classes	3	.85 (.10)	.79 (.14)	.82 (.12)
	4	.84 (.11)	.79 (.14)	.81 (.12)
	5	.80 (.13)	.71 (.16)	.76 (.14)
relative class size	balanced	.82 (.12)	.79 (.13)	.81 (.12)
	unbalanced	.84 (.11)	.73 (.17)	.79 (.14)
total		.83 (.12)	.76 (.15)	.80 (.13)

of .76) and restricted LCA (mean of .83).

The number of variables had the largest effect on average cluster recovery (partial $\eta^2 = .90$). As hypothesized, an increase in the number of variables from 5 to 10 increased the average cluster recovery substantially, from a mean of .68 to .91. The number of categories per variable also had a large effect on overall cluster recovery (partial $\eta^2 = .19$) in the hypothesized direction, with higher cluster recovery for 7 categories (mean of .82) than for 5 categories per variable (mean of .78). The effect of number of classes was also large (partial $\eta^2 = .36$): Mean overall cluster recovery decreased when there were more classes underlying the data, as was hypothesized (means of .82, .81, and .76 for the respective 3-, 4-, and 5-class populations).

The effects of number of categories and number of classes described above were similar for restricted LCA and restricted GROUPALS. Large differential effects for relative class size (partial $\eta^2 = .43$) and for number of variables (partial $\eta^2 = .22$) were found. However, there was also a large three-way interaction effect of these features with clustering technique (partial $\eta^2 = .19$). Figure 1 displays this effect graphically. Recovery of restricted LCA is unaffected by the relative sizes of the classes, both with 5 and 10 variables. However, as hypothesized, restricted GROUPALS performed worse when classes were unequally sized. With 5 variables, mean cluster recovery decreased from .68 to .59, while it only decreased from .91 to .88 with 10 variables. Apparently, including more parallel variables diminished the negative effect of unbalanced class sizes in restricted GROUPALS. So, with unequally sized classes, cluster recovery of restricted GROUPALS was poor with only 5 variables, but it was good with 10 variables.

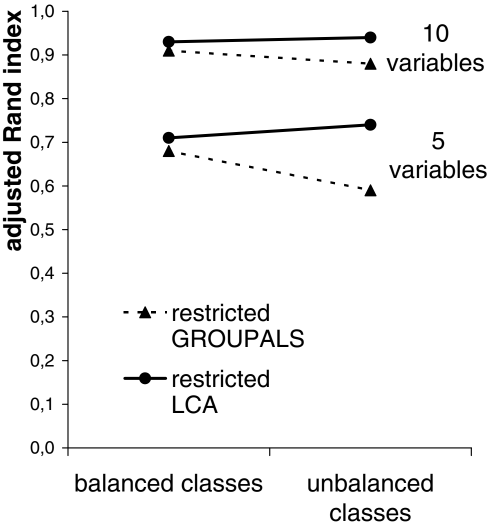
We studied whether restricted GROUPALS indeed found classes of roughly equal size as was stated before. We compared the sizes of the clusters obtained by restricted GROUPALS with the 'true' sizes of the clusters in the simulated data sets. For data with 3 classes, 5 variables with each 5 categories, mean true class sizes were .14, .29, and .57. Restricted GROUPALS resulted in mean class sizes of .21, .33, and .47, respectively. So, restricted GROUPALS indeed resulted in class sizes that were more equal than the

Table 3: Test statistics of effects on cluster recovery in Simulation Study I

effect	*SS* (type III)	*df*	*F*	partial η^2
between part of design				
no. of variables (VARS)	334.49	1	113364.6	.90
no. of categories (CAT)	8.46	1	2865.7	.19
no. of classes (CLASS)	20.08	2	3402.5	.36
rel. class size (RELCLASS)	2.11	1	715.2	.06
error	35.34	11976		
total	407.78			
within part of design				
clustering technique (CLUS)	26.55	1	24576.7	.67
CLUS x VARS	3.68	1	3407.2	.22
CLUS x CAT	1.15	1	1068.3	.08
CLUS x CLASS	1.02	2	471.5	.07
CLUS x RELCLASS	9.69	1	8966.6	.43
CLUS x VARS x RELCLASS	3.11	1	2886.3	.19
error (CLUS)	12.94	11976		
total	59.75			

Note. All effects significant with $p < .001$

Figure 1: Three-way interaction effect of relative class size, number of variables and clustering technique on cluster recovery



true class sizes. Class sizes estimates by restricted LCA were unbiased.

Finally, sample size was varied. For data with 5 variables with each 5 categories, additional simulations were done with sample size $N = 100$ and $N = 50$. Relative class size was varied. Again, restricted GROUPALS was negatively affected by unequally sized classes (partial $\eta^2 = .20$), decreasing mean recovery from .70 to .58, while restricted LCA was unaffected with mean recovery of .70 in both instances. Recovery decreased with decreasing sample sizes, yielding means of .74, .70, and .65 for restricted LCA and .65, .64, and .62 for restricted GROUPALS for $N = 300$, 100, and 50, respectively. However, this effect was not large, neither on average recovery (partial $\eta^2 = .06$), nor differentially (partial $\eta^2 = .06$). So, with sample sizes as small as 50, moderate cluster recovery can be obtained with restricted LCA and also with restricted GROUPALS, as long as the classes are not too unbalanced in size.

6.4.4 Conclusions

The following conclusions were drawn from this simulation study. Overall cluster recovery was good, and restricted GROUPALS approximated recovery by restricted LCA. For both techniques, the presence of more classes had a negative effect on cluster recovery, while increasing the number of categories per variable positively affected cluster recovery. Relative class size was predominantly a factor of importance for restricted GROUPALS, where unbalanced classes resulted in lower cluster recovery, but it did not affect cluster recovery for restricted LCA. Finally, increasing the number of variables had the largest positive effect on cluster recovery for both techniques, and it also seemed to alleviate the negative effect of unbalanced class size in restricted GROUPALS. Smaller samples did not deteriorate recovery.

6.5 Simulation Study II

By using artificial model parameters, the first simulation study may not be very realistic. Therefore, we carried out a second simulation study, based on results of an empirical analysis. Specifically, we analyzed a data set containing the strategies 574 pupils used to solve 5 math items on written division. The data came from a large national Dutch assessment carried out in 1997 (Janssen, Van der Schoot, Hemker, & Verhelst, 1999). The observed solution strategies were classified according to a coding system (Hickendorff, Heiser, Van Putten, & Verhelst, 2007; Rademakers, Van Putten, Beishuizen, & Janssen, 2004). Six different solution strategies were distinguished. Students could apply the *Traditional Algorithm* for long division, or apply the *Realistic Algorithm*, or solve the item in a more *Problem Solving*-like way. Furthermore, it was often observed that students answered an item without writing anything down: *No Working*. Finally, a solution strategy could be *Unclear*, or an item could be *Skipped*.

The six categories could also be collapsed to four. Problem solving approaches and the Realistic algorithm could also be classified as *Realistic* strategies, while Unclear strategies and Skipped items could be taken together in a remainder category *Other*. Frequency distributions of both the 4-category and the 6-category classification are in Table 4. On average, items were solved one third of the time by the traditional algorithm, and one third of the time without any written working. One fifth of the time items were solved by realistic strategies, mainly consisting of the realistic algorithm. The remainder category consisted mostly of skipping the item.

Table 4: Proportions of used solution strategies

strategy	item 1	item 2	item 3	item 4	item 5	average
Traditional Algorithm (T)	.28	.31	.37	.34	.31	.32
Realistic Strategies (R)	.18	.22	.20	.21	.17	.20
Realistic Algorithm (RA)	*.13*	*.16*	*.18*	*.16*	*.16*	*.16*
Problem Solving (PS)	*.05*	*.06*	*.02*	*.05*	*.01*	*.04*
No Working (NW)	.28	.41	.33	.26	.39	.34
Other (O)	.25	.06	.10	.19	.13	.15
Unclear (U)	*.05*	*.01*	*.01*	*.04*	*.00*	*.02*
Skipped (S)	*.21*	*.05*	*.09*	*.15*	*.13*	*.12*

Table 5: Conditional and class probabilities of Simulation Study II

	4 strategies					6 strategies						
class	T	R	NW	O	class size	T	RA	PS	NW	U	S	class size
α	.04	.05	.65	.26	*.34*	.04	.04	.02	.65	.01	.24	*.34*
β	.01	.66	.22	.11	*.26*	.01	.57	.09	.22	.03	.08	*.25*
γ	.75	.03	.15	.08	*.41*	.74	.00	.03	.15	.02	.05	*.41*
α	.04	.07	.29	.60	*.12*	.04	.08	.02	.28	.04	.54	*.14*
β	.08	.04	.80	.08	*.25*	.08	.03	.02	.80	.00	.07	*.24*
γ	.01	.67	.22	.09	*.25*	.01	.59	.09	.23	.02	.06	*.24*
δ	.77	.03	.12	.08	*.38*	.77	.00	.03	.13	.02	.05	*.38*
α	.01	.68	.22	.09	*.24*	.01	.62	.06	.23	.02	.06	*.22*
β	.50	.08	.38	.04	*.18*	.32	.08	.27	.26	.07	.00	*.06*
γ	.85	.01	.05	.09	*.26*	.79	.00	.01	.12	.02	.06	*.35*
δ	.02	.05	.84	.10	*.20*	.07	.03	.01	.81	.00	.08	*.24*
ϵ	.05	.07	.27	.62	*.12*	.04	.07	.03	.28	.03	.55	*.13*

6.5.1 Design

We varied the number of categories per variable, either 6 or the collapsed 4. Furthermore, we chose different numbers of classes: 3, 4, or 5.

6.5.2 Setting the Model Parameters

The empirical data matrix of 574 students by 5 variables with each either 4 or 6 categories was analyzed with restricted LCA. The conditional and class probabilities of 3-, 4-, and 5-class solutions were saved, and are displayed in Table 5. They served as the model parameters from which data were simulated according to the LC model, with $N = 574$. So, in total 6 different cells were included in the design. Again, each cell contained 500 replicated data sets. Each data set was clustered twice: once by restricted LCA and once by restricted GROUPALS.

6.5.3 Results

Means and standard deviations of cluster recovery of the 6 cells are reported in Table 6. Trends were partly similar as before. Cluster recovery by restricted LCA was much higher

Table 6: Means and standard deviations (in parentheses) of the adjusted Rand index in Simulation Study II

number of categories	number of classes	restricted LCA	restricted GROUPALS
4	3	.91 (.02)	.82 (.04)
	4	.88 (.02)	.82 (.03)
	5	.74 (.04)	.66 (.05)
6	3	.91 (.02)	.85 (.03)
	4	.87 (.02)	.83 (.03)
	5	.84 (.03)	.77 (.03)

than that of restricted GROUPALS (partial $\eta^2 = .76$). Furthermore, with more classes recovery decreased (partial $\eta^2 = .81$), while when variables had more categories recovery increased (partial $\eta^2 = .40$). However, there was a large interaction effect between number of classes and number of categories (partial $\eta^2 = .47$), caused by the recovery being only moderate with 5 classes and 4 categories. It seems that including more classes than categories deteriorated recovery, both for restricted LCA and for restricted GROUPALS. Finally, there was a large interaction effect between clustering technique and the number of classes (partial $\eta^2 = .12$). With 4 classes, recovery by restricted GROUPALS was higher than predicted by the main negative effect of an increasing number of classes. However, these results are hard to interpret, because class sizes and conditional probabilities also varied between the samples with 3, 4, and 5 classes.

6.5.4 Conclusions

Also on data simulated on model parameters based on results from empirical analyses, recovery of cluster membership was good to excellent for restricted LCA. Restricted GROUPALS approximated this optimal performance, but not very well. Only with data with more classes than categories, recovery was low, especially for restricted GROUPALS.

7. Discussion

Two techniques to obtain a partitioning of individuals measured on several replicated variables of categorical measurement level were discussed. Those were latent class analysis, with equality restrictions on the conditional probabilities, and a newly developed algorithm, GROUPALS with equality restrictions on the category quantifications. In two simulation studies, data were simulated according to the latent class model. In the first study, artificial model parameters were chosen to make comparisons across design factors possible. Results from an empirical analysis served as model parameters for the second study, to increase external validity.

Restricted LCA obtained moderate to excellent clustering results, dependent on several features of the data. Restricted GROUPALS approximated performance of restricted LCA reasonably well. The main exception was when the underlying classes are unequally sized. However, results from the second simulation study show good recovery with class sizes based on empirical results.

In the following, limitations of the simulation study are discussed, followed by a discussion of unresolved issues in restricted GROUPALS and in restricted LCA. Finally, some recommendations for practical applications are made.

7.1 Limitations of the Simulation Study

As for any simulation study, the main limitation of the present study is the problem of generalizability: Are the results only valid for the types of data structures in the present artificial data or can they be generalized to other data structures? The present types of data structures are defined by choices made in the generation of the data and choices made about the systematically varied data features.

Firstly, the choice was made to generate the data according to the latent class model. In this way, we were able to assess the most optimal performance by the recovery of restricted LCA, and compare performance of restricted GROUPALS to that performance.

In Simulation Study I, an important decision in the data generation has been to let the classes lie on a simplex, which is similar to what De Craen et al. (2006) did. This was mainly done to make the comparison between different cells of the design more justifiable. It was implicitly assumed that all configurations meeting this simplex-requirement would yield similar cluster recovery results. To test this assumption, in additional simulations, for one cell of the design two more configurations lying on a simplex were derived also with a distance between all classes of $\sqrt{.45}$. Contrary to our expectations, the specific configuration did affect cluster recovery, but moreover, it affected cluster recovery by restricted GROUPALS and by restricted LCA in opposite directions. So, this effect could have serious implications for the results of the present simulation study, since quite arbitrarily made choices in the determination of the conditional probabilities could have affected the cluster recovery of both techniques, influencing especially the effects for the number of classes and the number of categories.

In addition, in practical situations all classes do not need to be equally (dis)similar from each other (which is the case if classes lie on a simplex). Therefore, we included Simulation Study II, in which model parameters were derived from empirical analyses.

Another limitation is that the dependent variable of this study was bounded by 0 and 1, so a ceiling effect probably was present. This ceiling may have partially caused the decrease in variability and the apparent diminishing negative effect of unequal class sizes in restricted GROUPALS when the data consisted of more variables. However, logarithmic or square-root transformations of the adjusted Rand index did not change the results.

7.2 Issues in Restricted LCA

In the simulation study, the only aspect of LCA studied was the obtained classification based on estimated posterior probabilities, used for the computation of the dependent variable cluster recovery. Other parameters such as fit, estimated (conditional) probabilities and identification of the model were not studied, thereby limiting the scope of attention. One issue certainly deserving further study is the frequency of occurrence of local optima and whether there are factors influencing their prevalence. In our study it seemed that when the number of classes increased, the prevalence of a local optimal solution was larger. Indeed, Formann (2003) already argued that with more classes the likelihood surface is likely to become multimodal, so this could result in locally optimal solutions.

7.3 Issues in Restricted GROUPALS

As yet, the only criterion available to choose the best solution in GROUPALS (in the absence of external information on the clustering structure) is the fit of the solution. However, the solution with the best fit need not be the optimal solution with respect to cluster recovery, and the relation between fit and recovery might not even be positive.

To study this phenomenon, 200 solutions were obtained for a single sample, by starting with different initialization values. For all 200 solutions the cluster recoveries were computed, instead of only computing the cluster recovery of the solution with the best fit. This procedure was repeated for 500 samples, and in most samples it resulted in a small clump of analyses with a quite low fit and a quite low cluster recovery, and another large clump with a high fit and high recovery. Hence there appears to be a positive relation between fit and cluster recovery. However, within the clumps of solutions with high fit and recovery, there was no relation between fit and cluster recovery. So, by choosing the solution with the best fit, it was probably only assured that the worst solutions were filtered out, but not always the solution with the best cluster recovery was obtained. More generally, criterion-based clustering techniques would benefit from including some strategy that focuses on performance in future observations, such as cross-validation.

The negative effect of classes that are unbalanced in size on cluster recovery by restricted GROUPALS directs attention to the incorporated K-means algorithm. It would be interesting to redefine the GROUPALS loss function to incorporate some other clustering technique that can handle classes of unequal size and of other shape, such as model-based clustering (e.g., Fraley & Raftery, 2002) or fuzzy clustering.

A final discussion point in GROUPALS is the argument that the simultaneous scaling and clustering should be more sensitive to clustering structure in the data than the sequential approach (Van Buuren & Heiser, 1989; Vichi & Kiers, 2001). However, from some preliminary analyses (not discussed here further), promising results were obtained from a two-step procedures of correspondence analysis on the frequency table F followed by a clustering algorithm. So, discrediting a two-step procedure in advance may not be warranted. It would be interesting for future research to study such sequential approaches in more detail, especially in comparison with the simultaneous scaling and clustering in (restricted) GROUPALS.

7.4 Recommendations

Finally, some recommendations for practical applications are made. Firstly, a researcher might check that the variables in the study are replications in measuring some construct, by conducting a multiple correspondence analysis. It can be explored whether equivalent categories indeed obtain similar quantifications, or the decrease in fit by imposing equality restrictions on the category quantifications in multiple correspondence analysis may be assessed. It is also possible to test equality restrictions in LCA directly by computing the Likelihood Ratio statistic of models with and without equality constraints. Furthermore, imposing equality restrictions may be necessary to avoid identification problems when fitting LC models to data with many variables and/or many categories per variable.

When the researcher is convinced that indeed the variables are parallel indicators, and he or she wants to derive a partitioning of the individuals in different classes, restricted LCA is the technique of choice. Restricted GROUPALS can be used as an approximation, especially if one is interested in scaling of the variables in addition to obtaining a clustering of the individuals, provided that the classes are not expected to be extremely different in size.

References

Arabie, P., & Hubert, L.J. (1996). An overview of combinatorial data analysis. In P. Arabie, L.J. Hubert & G. De Soete (Eds.), *Clustering and classification* (p. 5–64). Singapore: World Scientific Publishing.

Chaturvedi, A., Green, P.E., & Carroll, J.D. (2001). K-modes clustering. *Journal of Classification*, *18*, 36–55.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L.M., Fidler, P.L., Wugalter, S.E., & Long, J.D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*, 375–389.

De Craen, S., Commandeur, J.J.F., Frank, L.E., & Heiser, W.J. (2006). Effects of group size and lack of sphericity on the recovery of clusters in K-means cluster analysis. *Multivariate Behavioral Research*, *41*, 127–145.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Formann, A.K. (2003). Latent class model diagnosis from a frequentist point of view. *Biometrics*, *59*, 189–196.

Fraley, C., & Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.

Gifi, A. (1990). *Nonlinear multivariate data analysis*. Chichester, England: John Wiley & Sons.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Goodman, L.A. (2002). Latent class analysis: The empirical study of latent types, latent variables and latent structures. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied latent class analysis* (p. 3–55). Cambridge: Cambridge University Press.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M.J. (1988). Clustering the rows and columns of a contingency table. *Journal of Classification*, *5*, 39–51.

Heiser, W.J. (1981). *Unfolding analysis of proximity data*. Doctoral dissertation, Leiden University, Leiden, The Netherlands.

Hickendorff, M., Heiser, W.J., Van Putten, C.M., & Verhelst, N.D. (2007). Solution strategies and achievement in Dutch written arithmetic: Latent variable modeling of change. *Manuscript submitted for publication.*

Hosenfeld, B., Van der Maas, H.L.J., & Van den Boom, D.C. (1997). Detecting bimodality in the analogical reasoning performance of elementary schoolchildren. *International Journal of Behavioral Development*, *20*, 529–547.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Hwang, H.S., Dillon, W.R., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, *71*, 161–171.

Janssen, J., Van der Schoot, F., Hemker, B., & Verhelst, N.D. (1999). *Balans van het rekenwiskundeonderwijs aan het einde van de basisschool 3* [Third assessment of mathematics education at the end of primary school]. Arnhem, The Netherlands: CITO.

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis.* New York: Houghton-Mifflin.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In L.M. Le Cam & J. Neyman (Eds.), *Proceedings of 5-th Berkeley symposium on mathematical statistics and probability* (Vol. 1, p. 281–297). Berkeley, CA: University of California Press.

McCutcheon, A.L. (1987). *Latent class analysis.* Beverly Hills, CA: Sage Publications.

Mooijaart, A., & Van der Heijden, P.G.M. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika*, *57*, 261–269.

Rademakers, G., Van Putten, C.M., Beishuizen, M., & Janssen, J. (2004). Traditionele en realistische algoritmes bij het oplossen van deelsommen in groep 8: een nadere analyse van PPON-materiaal uit 1997 [Traditional and realistic algorithms for solving division problems in grade 6: A further analysis of the material of the 1997-assessment]. *Rekenwiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, *23*, 3–7.

Steinley, D. (2003). *K*-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294–304.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*, 386–396.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using mutivariate statistics* (4th ed.). New York: Allyn and Bacon.

Thomas, H., & Hettmansperger, T.P. (2001). Modelling change in cognitive understanding with finite mixtures. *Applied Statistics*, *50*, 435–448.

Van Buuren, S., & De Leeuw, J. (1992). Equality constraints in multiple correspondence analysis. *Multivariate Behavioral Research*, *27*, 567–583.

Van Buuren, S., & Heiser, W.J. (1989). Clustering N objects into K groups under optimal scaling of variables. *Psychometrika*, *54*, 699–706.

Van Putten, C.M., Van den Brom-Snijders, P.A., & Beishuizen, M. (2005). Progressive mathematization of long division strategies in Dutch primary schools. *Journal for Research in Mathematics Education*, *36*, 44–73.

Vermunt, J.K. (1997). *LEM 1.0: A general program for the analysis of categorical data.* Tilburg, The Netherlands: Tilburg University.

Vermunt, J.K., & Magidson, J. (2002). Latent class cluster analysis. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied latent class analysis* (p. 89–106). Cambridge, England: Cambridge University Press.

Vichi, M., & Kiers, H.A.L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*, 49–64.

(Received July 12 2007, Revised March 4 2008)