Measurement and Research Department Reports 2005-3

Clustering Nominal Data with Equivalent Categories: a Simulation Study Comparing Restricted GROUPALS and Restricted Latent Class Analysis

Marian Hickendorff

cito groep

Measurement and Research Department Reports 2005-3

Clustering Nominal Data with Equivalent Categories: a Simulation Study Comparing Restricted GROUPALS and Restricted Latent Class Analysis

Masters Thesis in Psychology

Marian Hickendorff

Division of Methodology and Psychometrics
Leiden University, and
Psychometric Research Center
CITO, Dutch National Institute of Educational Measurement

Supervisors
Prof. Dr. Willem J. Heiser
Dr. Cornelis M. van Putten
Dr. Norman D. Verhelst

Cito
Arnhem, 2005

TABLE OF CONTENTS

FOREWORD

This study has been carried out as a thesis for the Masters degree of Psychology at Leiden University in the division of Methodology and Psychometrics. It was supported by a grant for Masters students in Psychometrics from CITO, the National Institute of Educational Measurement. I would like to thank CITO for giving me the opportunity to work there and the flexibility I was allowed to work with, and I thank all my colleagues there for their support and helpfulness.

I would like to use this moment to thank to my supervisors. I thank Kees van Putten for starting this thesis up and for his continuing support during the process, which especially during the at times struggling start of this study was indispensable. I thank Willem Heiser for freeing up time for me in his sabbatical year and for the input of his impressive psychometric knowledge, without which this study would have been far less sophisticated. Finally, I thank Norman Verhelst for his unceasing enthusiasm, even though he had to deal with a study already set up before he joined in, for his flexibility, his willingness to help me at all times, his refreshing ideas and criticisms, his psychometric input and above all for his time and confidence he was willing to invest in me.

Marian Hickendorff

Leiden, May 18, 2005

ABSTRACT

This study discusses methods for clustering data of nominal measurement level, where the categories of the variables are equivalent: the variables are considered as parallel indicators. Two techniques were compared on their cluster recovery in the analysis of simulated data sets with known cluster structure, by means of the adjusted Rand index.

The first technique was GROUPALS, an algorithm for the simultaneous scaling (by homogeneity analysis) and clustering of categorical variables. To account for equivalent categories, equality constraints of the category quantifications for the same categories of the different variables were incorporated in the GROUPALS algorithm, resulting in a new technique. The second technique was latent class analysis, with the extra restriction to account for parallel indicators that the conditional probabilities were equal across the variables.

Restricted LCA obtained higher cluster recovery than restricted GROUPALS. For both techniques, increasing the number of variables and the number of categories per variable positively affected cluster recovery, while the presence of more classes negatively affected cluster recovery. These effects were more pronounced for restricted GROUPALS. Relative class size was predominantly a factor of importance for restricted GROUPALS, where unbalanced classes negatively affected cluster recovery, but it did not affect cluster recovery for restricted LCA much. Finally, increasing the number of variables seemed to alleviate the negative effect of more underlying classes for both techniques and the negative effect of unbalanced class size in restricted GROUPALS.

1. INTRODUCTION

This study concerns techniques for the clustering of categorical data, where the categories of all the variables are equivalent: the variables are parallel indicators. In this introduction, first the concept of clustering is discussed, followed by a description of the two special characteristics of the data considered in this study. Next, techniques that are suitable to cluster this type of data are described: GROUPALS and latent class analysis, and then some adjustments in the form of restrictions to both techniques are discussed. Finally, the purpose of this study and the research questions are specified.

1.1 Clustering

Clustering is a form of data analysis, and can be characterized as those methods that are concerned with the identification of homogeneous groups of objects, based on data available (Arabie & Hubert, 1996). The purpose of clustering methods is to identify the possible group structure, such that objects in the same group are similar in some respect – the desideratum of *internal cohesion* – and different from objects in other groups – the desideratum of *external isolation* (Gordon, 1999). It should be noted that the terms 'clustering' and 'classification' are used here interchangeably and refer to the situation in which the classes or clusters are unknown at the start of the investigation: the number of classes, their defining characteristics and their constituent objects need to be determined by the clustering or classification method (Gordon, 1999). These methods should be clearly distinguished from methods that are concerned with already-defined classes, such as discriminant analysis (Tabachnick & Fidell, 2001), forced classification (Nishisato, 1984) and pattern recognition.

Two classical approaches to the algorithm for identifying the clusters exist (Van Os, 2000). First there is hierarchical cluster analysis, which can be either agglomerative or, less commonly, divisive. Agglomerative hierarchical cluster analysis starts with all objects in separate classes and successively joins groups that exist at previous levels in a step-by-step procedure. Divisive hierarchical cluster analysis works in the opposite direction, starting with all objects in one group and proceeding by successively splitting classes step-by-step.

In contrast there are partitioning methods, which seek one single partition of the objects in mutually exclusive and exhausting subsets (Van Os, 2000). Typically, a measure for the 'goodness' of any proposed partition is defined, and the purpose of the partition methods is to find the partition that is optimal with respect to this measure. These methods are therefore also referred to as optimization techniques for clustering (Everitt & Dunn, 2001). One of the more popular of these procedures is the *K*-means algorithm, which iteratively relocates objects between classes, until no further improvement of the measure to be optimized can be obtained.

The types of input data for both of the two classical approaches to cluster analysis are typically characterized by 'ways' and 'modes'. Ways refer to the dimensions of the data table: a table with rows and columns is a two-way table. How many modes a data matrix has, is concerned with how many sets of entities it refers to. If the ways of the data matrix correspond to the same set of entities, such as with similarities or distances between pairs of objects, the data are one-mode. Two-mode data involve two sets of entities, such as objects and variables (Arabie & Hubert, 1996). Many methods of clustering require two-way, one-mode data (either similarities or Euclidean distances between the objects, an N objects by N objects matrix). However, the data as collected by researchers often consist of several variables measured on the objects (an N objects by m variables matrix) and as such are two-way, two-mode. As a preprocessing step prior to the classical approaches to cluster analysis, some type of conversion of the data from two- to one-mode is necessary, and several techniques have been proposed (Arabie & Hubert, 1996). The measurement level of the variables is an important consideration for the type of conversion. Especially the case of clustering categorical data raises problems to this conversion issue, and clustering of categorical data is what is studied here. Only techniques that obtain a single partition of the objects are discussed (as opposed to hierarchical methods, which are beyond the scope of this paper).

Besides the classical approaches to cluster analysis described, there are recent developments in clustering in the form of mixture models or probabilistic clustering (Arabie & Hubert, 1996). When all variables are categorical, these models reduce to latent class analysis (e.g. Bacher, 2000; Everitt & Dunn, 2001; Bock, 1996). These probabilistic or mixture models also have the purpose of deriving a partitioning of the objects in classes. An important difference between the classical

clustering methods and probabilistic or mixture clustering techniques is that the latter are model-based, meaning 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. Vermunt and Magidson (2002) pose several advantages of the model-based approach over the standard approaches. Formal criteria can be used to test the model features, and the validity of restrictions, that can be imposed on parameters, can be tested statistically. Furthermore, the clustering is probabilistic instead of deterministic, so the uncertainty of the classification is also estimated.

Nevertheless, the log-likelihood function that has to be optimized in LCA may be very similar to the criterion optimized in certain classical partitioning procedures such as *K*-means (Vermunt & Magidson, 2002). So, the difference between model-based clustering techniques and the classical partitioning techniques may be more an issue of difference in theoretical point of view, than of difference in practical point of view.

1.2 Type of data sets considered

The input data considered in this study are two-way, two-mode data (objects measured on several variables), because this is what will be encountered in practical situations most often. This study only focused on data sets with two special features. Data with these features were encountered in research on mathematics education (Van Putten, Van Den Brom-Snijders, & Beishuizen, 2005). From a sample of pupils, the strategies they used to solve several division exercises were coded in a category system. There were reasons to suspect several classes of pupils with their own characteristics of strategy use, so information on suitable clustering techniques was needed. The data had the following two special features:

1) All variables are of nominal measurement level.

This means that all variables are categorical, where for example the categories code the different strategies pupils apply to mathematics exercises.

2) All variables are parallel indicators, meaning that they have 'equivalent categories'.

All variables have an equal number of categories, and for example category 3 codes the same strategy for all variables. So, applying strategy 3 on exercise (variable) 4 has the same meaning as applying strategy 3 on exercise 7. The variables can be viewed as replications.

First, techniques that are suitable to cluster categorical data (the first characteristic) are explored, followed by a discussion how these techniques can cope with variables with equivalent categories (the second characteristic), by means of equality restrictions.

1.3 Techniques for partitioning categorical data

Partitioning of categorical data is not as straightforward as clustering of numerical data. Since the numerical values of categorical data are meaningless (these are just arbitrary codes for the categories), and partitioning procedures such as *K*-means are based on numerical input, categorical data need special attention.

Chaturvedi, Green, and Carroll (2001) sum up five techniques that are commonly used for finding clusters in categorical data. The first two techniques dummy code the categorical variables, and on the dummy coded data either the intersubject distances are computed and on these distances a hierarchical clustering method is applied, or *K*-means is applied to the dummy coded data. The authors note that the former method has the drawbacks that a distance measure should be selected and that hierarchical clustering procedures do not optimize an explicit measure of fit, and that the latter technique is inappropriate because the *K*-means algorithm minimizes an ordinary least-squares function which is not valid for categorical data, and means are not appropriate measures of central tendency in categorical data.

Thirdly, the authors make reference to the Ditto Algorithm of Hartigan (1975) but claim that the algorithm does not even guarantee locally optimal solutions. Fourthly, latent class procedures are an option for clustering categorical data. These techniques are theoretically sound but can become computationally intense and they rely on assumptions of local independence and certain parametric assumptions about the nature of the data.

Finally, as a way to overcome the problem that the codes for the categories are meaningless as numerical values, it is possible to first use correspondence analysis as a dimension reduction technique to derive numerical values for the categories, and then use *K*-means on these derived

spatial coordinates. This so-called tandem analysis may be inappropriate, as noted by several authors (Chaturvedi et al., 2001; Vichi & Kiers, 2000). That is, 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 en Heiser (1989) proposed a method called GROUPALS, in which the scaling of the variables and the clustering are done simultaneously, so that the solution is optimal to both criteria at the same time. Such a model is also proposed for numerical data, so-called factorial *K*-means (Vichi & Kiers, 2000).

Techniques that are appropriate to cluster categorical data are therefore the GROUPALS technique proposed by Van Buuren and Heiser (1989), and latent class analysis (e.g. McCutcheon, 1987; Hagenaars & McCutcheon, 2002). A comparison of the clustering performance of these techniques, that is, the ability to recover a known cluster structure in a data set, is the purpose of this paper. Hereafter, both latent class analysis and GROUPALS are described. For both techniques the basic, unrestricted model is discussed, followed by an exploration of the restrictions that need to be imposed on the model to account for the equivalent categories of all the variables in the data set. In the following, let H be the data matrix of the form N objects by m categorical variables each with l_j $(j = 1, \ldots, m)$ categories, and let K denote the number of classes or clusters the objects belong to.

1.4 GROUPALS

1.4.1 Basic concepts of GROUPALS[1]

The first technique in this study is GROUPALS, a clustering method proposed by Van Buuren and Heiser (1989). It has the purpose of reducing many variables with mixed measurement level to one (cluster allocation) variable with K categories. As already noted, the rationale of the technique is the simultaneous *clustering* of the objects by a *K*-means procedure and *scaling* of the

[1] The present discussion of GROUPALS and restricted GROUPALS is at times very similar to the discussion by Van Buuren and Heiser (1989), especially when the steps of the algorithm are described. Please note that this was done to present the sometimes complex discussion in a way that is comparable to the original discussion of GROUPALS.

objects by an optimal scaling technique. The desideratum of internal cohesion should be satisfied by the *K*-means algorithm of minimizing trace(*W*) with *W* the pooled-within group sum-of-squares matrix (Van Buuren, 1986). The external isolation desideratum is taken care of by optimal scaling which determines new variables with maximum variation by making specific linear combinations of the observed variables. In the case of categorical variables, optimal scaling is performed by a technique called multiple correspondence analysis, also called homogeneity analysis or HOMALS (HOMogeneity Analysis by Alternating Least Squares).

Optimal scaling is a technique to derive a transformation of the variables in a data set such that the correlations between the transformed variables are maximized (Gifi, 1990). These transformations can be viewed as quantifications in some predetermined number of dimensions *(p)* of the category codes, so that meaningful numerical values are obtained. In the following, define X as the $(N \times p)$ matrix of object scores with for each object quantifications in p dimensions, and define the m Y_j $(j = 1, \ldots, m)$ matrices of size $(l_j \times p)$ as containing the category quantifications for each of the m variables in p dimensions. The m G_j matrices $(N \times l_j)$ are indicator matrices for each of the m categorical variables. To estimate these optimal quantifications for objects and categories, the HOMALS loss function

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

should be minimized over X and the m Y_j matrices. This minimization can be carried out by an alternating least squares algorithm, usually with normalization $X'X = I$. The resulting solution contains the category quantifications and the object scores in p dimensions. This solution is optimal in the sense that the dimensions have maximum explained variance.

An important advantage in the context of clustering is that the meaningless arbitrary scores of the objects on the categorical variables are transformed to numerical values: the object scores. So, now the problem of data with arbitrary numbers is overcome, and it is possible to apply a partitioning method on the derived quantifications (the earlier mentioned two-step approach called tandem analysis).

However, Van Buuren and Heiser (1989) note that this optimization algorithm does not guarantee that the obtained scores in the dimensions are also optimal for identifying a possibly

existent clustering structure in the data. Therefore, they propose GROUPALS, in which an extra restriction on the loss-function for HOMALS is inserted: all objects in the same group should be at the same position (at the cluster mean) in the p-dimensional space. The advantage is that the dimension reduction and clustering are performed simultaneously, instead of sequentially as in the tandem analysis. This results in the following GROUPALS loss function

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

which is also optimized by an alternating least squares algorithm over G_c, Y_c and the m Y_j matrices.[2] G_c *(N* x *K)* is the indicator matrix for cluster allocation of the objects and Y_c is the *(K* x *p)* matrix of cluster points. Estimated are the cluster allocation of the objects, the positions of the clusters and the category quantifications.

Limitations of the technique are that it is likely to produce local optimal solutions and the tendency to produce spherical, equally sized clusters, inherent to the *K*-means algorithm.

1.4.2 GROUPALS with equality restrictions

If all variables in the data set have equivalent categories, the basic GROUPALS loss function can be adjusted. Defining the categories of all variables to be equivalent can be operationalized as 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 equal to the category quantifications of category 1 of variables *B* and *C*, for all dimensions of the solution.

What is required in this GROUPALS with equality restrictions on the category quantifications is optimal scaling of the objects and categories, under the restriction that objects in the same cluster are on the same position on the 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 this from now on called restricted GROUPALS is as follows:

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

$$\sigma(G_c; Y_c; Y) = \frac{1}{m}\sum_{j=1}^{m} tr(G_c Y_c - G_j Y)'(G_c Y_c - G_j Y) \quad (3)$$

For fixed G_c and Y_c, the loss function has to be minimized over Y. To do this, the partial derivative of (3) with respect to Y has to be derived and set equal to zero, to obtain an estimate of the quantifications Y. First, the loss-function (3) can be rewritten as follows:

$$\sigma(G_c; Y_c; Y) = tr(G_c Y_c)'(G_c Y_c) + \frac{1}{m}\sum_{j=1}^{m} tr(G_j Y)'(G_j Y) - \frac{2}{m}\sum_{j=1}^{m} tr(G_c Y_c)' G_j Y \quad (4)$$

Define $F = \sum_j G_j$; $D_j = G_j' G_j$ and $D = \sum_j D_j$. F is a frequency matrix of size N x l ($l_1 = \ldots l_m = l$), with for all objects the number of times they chose category 1, 2, 3... etc. 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 objects, and D is a diagonal matrix (l x l) with the frequencies of the categories, summed over all j variables. Now (4) can be simplified to

$$\sigma(G_c; Y_c; Y) = tr(G_c Y_c)'(G_c Y_c) + \frac{1}{m} tr(Y' D Y) - \frac{2}{m} tr(G_c Y_c)' F Y \quad (5)$$

The partial derivative of (5) with respect to Y is

$$\frac{\partial \sigma(G_c; Y_c; Y)}{\partial Y} = \frac{2}{m} D Y - \frac{2}{m} F' G_c Y_c \quad (6)$$

and setting this equal to zero and solving for Y gives the following equation to compute the optimal quantifications Y:

$$\hat{Y} = D^{-1} F' G_c Y_c \quad (7)$$

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

It should be noted that this is the same formula that is 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 HOMALS loss function (without cluster restrictions on the object scores) with

equality restrictions on the category quantifications is equivalent to performing correspondence analysis on the frequency matrix $F = \sum_j G_j$. So the quantification step in restricted GROUPALS is the same as the quantification step in correspondence analysis on the frequency matrix F, if correspondence analysis is performed by an alternating least squares algorithm.

If Y is fixed, the loss has to be minimized over G_c and Y_c. Letting $Z = \frac{1}{m}\sum_j G_j Y$ (unrestricted object scores computed form the category quantifications Y), and inserting the identity $G_c Y_c = Z - (Z - G_c Y_c)$ into (3), the loss function can also be split into additive components as follows:

$$\sigma(G_c; Y_c; Y) = \frac{1}{m}\sum_{j=1}^{m} tr(Z - G_j Y)'(Z - G_j Y) + tr(Z - G_c Y_c)'(Z - G_c Y_c) \qquad (8)$$

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. This problem is known as the sum of squared distances (SSQD) clustering, and the SSQD criterion can be minimized by the iterative K-means algorithm. This results in the cluster allocation matrix G_c, and then the criterion is minimized by setting $Y_c := (G_c' G_c)^{-1} G_c' Z$ (the position of each cluster is the centroid of all objects belonging to that cluster).

Transfer of normalization

In order to prevent the algorithm from making X and Y zero, either the object scores X, or the category quantifications Y need to be normalized. In (unrestricted) HOMALS, it is conventional to use the normalization $X'X = I$. This works well in minimizing the unrestricted HOMALS loss function (1). However, in minimizing loss function (3) with restrictions on the object scores, this normalization results in two types of restrictions on the object scores: normalization and clustering restrictions, which leads to computational complications. Similarly, normalization of the category quantifications Y, by requiring $Y'DY = I$, is inconvenient.

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. Suppose there is some solution with normalization $X'X = I$, then nonsingular transformation matrices P and Q can be found such that $\sigma(X; Y) = \sigma(XP; YQ)$ with normalization $(YQ)'DYQ = I$, by using $P = K\Lambda$ and $Q = K\Lambda^{-1}$ from the eigenvalue decomposition $\frac{1}{m} Y'DY = K\Lambda^2 K'$. If this procedure is

applied twice in the algorithm, G_c and Y_c are estimated under normalization $Y'DY = I$ and Y under normalization $X'X = I$.

Algorithm restricted GROUPALS

The following steps constitute the algorithm of restricted GROUPALS. Note that this is very similar to the algorithm Van Buuren and Heiser (1989) describe for (unrestricted) GROUPALS. The only real difference lies in the quantification step. Further seemingly different algorithmic steps just arise from more efficient notation, made possible by the equality restriction on the category quantifications.

Step 1: Initialization

Set the number of clusters K and set the dimensionality of the solution p. Construct m indicator matrices G_j, and define $F = \sum_j G_j$; $D_j = G_j'G_j$ and $D = \sum_j D_j$. Set X^0 with orthonormalized, centered random numbers and set the indicator matrix G_c^0 with some initial partition. Set iteration counter $t = 1$.

Step 2: Quantification

Minimize the loss (4) over Y for a given X^{t-1}. As shown before, this can be done by setting: $Y^t = D^{-1}F'X^{t-1}$.

Step 3: Transfer of normalization to the quantifications

Define $T^t = \frac{1}{m} Y^{t\prime} D Y^t$ and compute the eigenvalue decomposition of $T = K\Lambda^2 K'$. Define $Z^t := \frac{1}{m} F Y^t K \Lambda^{-1}$.

Step 4: Estimation of cluster allocations

Minimize the SSQD criterion $tr(Z^t - G_c Y_c)'(Z^t - G_c Y_c)$, which is the second part of the loss function as written in (8), over G_c and Y_c, given Z^t and G_c^{t-1}, by means of the K-means algorithm. This results in G_c^t, and then set $Y_c^t := (G_c^{t\prime} G_c^t)^{-1} G_c^t Z^t$. Finally, define $X^{*t} := G_c^t Y_c^t$.

Step 5: Transfer of normalization to object scores

Compute the eigenvalue decomposition of $X^{*t} = L\Psi^2 L'$. Let $X^t := X^{*t} L\Psi^{-1}$.

Step 6: Convergence test

Compute the value of the loss function (3) and check whether the difference between the values at iterations t and t - 1 is smaller then some predetermined criterion value, or whether a maximum number of iterations has been reached. If so, stop; otherwise, set $t := t + 1$ and go to Step 2.

1.5 Latent class analysis

1.5.1 Basic concepts of latent class analysis

The second clustering technique in this study is latent class analysis (LCA). As already noted, LCA is actually a form of mixture or probability models for classification, where all data are categorical. The model postulates underlying probability functions generating the data, and these probability functions are in the case of categorical data usually assumed to be multinomial. The latent class model assumes an underlying latent categorical variable (coding the K latent classes) that can explain, or worded differently can explain away, the covariation between the observed or manifest variables (McCutcheon, 1987; Goodman, 2002). LCA can be seen as the categorical analogue to factor analysis. The purpose of LCA is to identify a set of mutually exclusive latent classes, and therefore the technique fits in the definition of a partitioning method.

The rationale of the technique is the axiom of local independence, meaning that, conditionally on the level of the latent class variable (named here T with K levels, coded α, β, γ etcetera), the probability functions are statistically independent (McCutcheon, 1987). This means that, in the hypothetical case of three observed variables (A, B and C) and one latent variable T that the latent class model can be expressed in a formula as the product of the latent class probabilities and the conditional probabilities as follows:

$$\pi_{ijl\kappa}^{ABCT} = \pi_{\kappa}^{T} \cdot \pi_{i|\kappa}^{A|T} \cdot \pi_{j|\kappa}^{B|T} \cdot \pi_{l|\kappa}^{C|T} \qquad (9)$$

$\pi_{ijl\kappa}^{ABCT}$ is the probability of an object scoring category i on variable A, category j on variable B, category l on variable C, and category κ on (latent class) variable T. This can be expressed as the product of the latent class probability π_{κ}^{T} which is the probability that an object falls in

latent class κ, and the conditional probabilities of scoring category i on variable A, conditional upon being in latent class κ ($\pi_{i|\kappa}^{A|T}$), of scoring category j on variable B, conditional upon being in latent class κ ($\pi_{j|\kappa}^{B|T}$) and of scoring category l on variable C, conditional upon being in latent class κ ($\pi_{l|\kappa}^{C|T}$). For example, the probability for an object to be in the first latent class and scoring category 2, 3 and 2 on variables A, B and C respectively, is the probability of being in latent class α, multiplied by the probability given that an object is in latent class α, of scoring category 2 on variable A, of scoring category 3 on variable B and of scoring category 2 on variable C.

Two sorts of model parameters are estimated: the latent class probabilities, corresponding to the class sizes, and the conditional probabilities, analogue to factor loadings in factor analysis. The latter are the probabilities of scoring a certain category on a certain variable, given the latent class an object is in, and can be helpful in identifying the characteristics of the latent classes.

The model is estimated by an iterative maximum likelihood procedure. Two commonly used algorithms are the expectation-maximization (EM) and the Newton-Raphson (NR) algorithms, both with their respective strong and weak points (for a further discussion, see McCutcheon, 2002). For both models, the functions to be optimized in parameter estimation suffer from local optima. Several criteria for model evaluation are available: the likelihood statistics χ^2 and L^2, that can test whether the model statistically fits the data, and information criteria AIC and BIC which penalize the likelihood criteria for increase in number of estimated parameters. These information criteria make it possible to compare different models, even if they are not nested.

Although in LCA a partitioning of objects is not estimated directly, one can derive a partitioning from the estimated parameters: the latent class probabilities and the conditional probabilities. From these estimates, it is possible to determine the posterior probability that an object, given its response pattern on the manifest variables, is in latent class κ, by means of:

$$\pi_{\kappa|ijl}^{T|ABC} = \frac{\pi_{\kappa}^{T} \cdot \pi_{i|\kappa}^{A|T} \cdot \pi_{j|\kappa}^{B|T} \cdot \pi_{l|\kappa}^{C|T}}{\sum_{\kappa} \pi_{\kappa}^{T} \cdot \pi_{i|\kappa}^{A|T} \cdot \pi_{j|\kappa}^{B|T} \cdot \pi_{l|\kappa}^{C|T}} \tag{10}$$

A conventional procedure is to ascribe an object to the latent class for which it has the highest posterior probability (modal assignment). This results in probabilistic classification, where it is also possible to asses the degree of uncertainty.

Limitations of LCA are the earlier mentioned occurrence of local optima in the estimation algorithms and the question of identification of the model. The log-likelihood function to be optimized can suffer from local minima to which the estimation algorithm converges, resulting in a locally optimal solution. Identification of the model is an issue, when too many variables and / or variables with too many categories are inserted in the model. Then, too many parameters have to be estimated given the data matrix (which is then said to be 'sparse') and the model is not identified (Collins, Fidler and Wugalter, 1996). Restrictions on parameters, either equality constraints or specific value constraints, can solve this identification problem. In the case of a sparse data matrix, another problem is that the likelihood statistics, to evaluate if the data fit the model, do not follow a χ^2-distribution, so these tests can not be trusted (Collins et al., 1996).

1.5.2 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 should be imposed. The case of parallel indicators in LCA means that the variables are assumed to measure the same construct, with the same error rate (McCutcheon, 2002). Technically, it boils down to restricting the conditional probabilities to be equal over the variables. In the hypothetical example of three parallel indicator variables A, B and C, this equality restriction is as follows: $\pi_{i|\kappa}^{A|T} = \pi_{i|\kappa}^{B|T} = \pi_{i|\kappa}^{C|T}$, for all categories $i = 1,...,l$ and for all classes κ. From now on, when the restricted LC model is mentioned, this refers to the LC model with equality restrictions on the conditional probabilities.

1.6 Theoretical comparison of LCA and GROUPALS

Although both LCA and GROUPALS are capable of clustering categorical data, these techniques approach the problem from very different perspectives. The main difference lies in the postulation of an underlying model in LCA, which is not the case in GROUPALS. This is not

only of theoretical interest, it has practical consequences too. It can be tested if the data significantly depart form the model estimated by LCA, and hence the model can be statistically rejected or accepted as fitting the data. This is not possible with GROUPALS, where 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.

The existence in LCA of information criteria (AIC and BIC) is also very convenient for determining the fit, since it allows for comparison of models that are not nested. For example, the comparison of the fit of a three-class with a four-class model on the basis of the likelihood statistics will be misleading, since more classes will result in a larger likelihood in any case. The information criteria 'penalize' the likelihood statistics for the increase in estimated parameters and therefore also take parsimony of the model into account. These criteria can be used in comparing the fit between the models. No such 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.

However, a drawback of the underlying model in LCA is that the number of parameters that have to be estimated increases very rapidly as the number of variables and / or the number of categories increases. This makes LCA computationally intense and can also result in identification problems.

Interpreting the solutions from GROUPALS and from LCA also occurs on very different grounds. In LCA, the estimated parameters, the latent class probabilities and the 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.

A disadvantage of both LCA and GROUPALS 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. In both cases, the only 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. Again, LCA has the advantage of criteria for these comparisons.

Another problem of both methods is the occurrence of local optima in the optimization algorithms. To deal with this, several starting configurations should be tried and the solution with the best fit should be interpreted.

1.7 Purpose of this study and research questions

Generally speaking, the goal of the present study is a comparison of the cluster recovery abilities of restricted latent class analysis and restricted GROUPALS in the partitioning of objects, when the data available are nominal variables with equivalent categories. In a Monte Carlo simulation study, data sets with known cluster structure were generated, so that the two methods could be compared on their cluster recovery capabilities. Several parameters were varied: the number of classes, the relative class sizes, the number of categories per variable and the number of variables. The research questions were therefore:

Main effect of repeated measure variable portioning technique

1) Overall, does LCA or GROUPALS have the highest cluster recovery?

Main effects of data aspects

2) Overall, what is the effect of the number of variables on cluster recovery?

3) Overall, what is the effect of the number of classes on cluster recovery?

4) Overall, what is the effect of the number of categories per variable on cluster recovery?

5) Overall, what is the effect of the relative class size on cluster recovery?

Interaction effects of data aspects with partitioning technique

6) What is the effect of number of variables on the comparison of LCA and GROUPALS?

7) What is the effect of number of classes on the comparison of LCA and GROUPALS?

8) What is the effect of number of categories per variable on the comparison of LCA and GROUPALS?

9) What is the effect of relative class size on the comparison of LCA and GROUPALS?

Further relevant effects

10) Are there further significant and practically important effects on cluster recovery?

1.8 Hypotheses

In their study comparing LCA with another clustering method called *K*-modes clustering, Chaturvedi et al. (2001) find that, for both LCA and *K*-modes, the number of classes is negatively related to cluster recovery, while the number of variables and the number of categories per variable are positively related to cluster recovery. Therefore, the hypotheses for the main effects of these independent variables are analogous. It is hypothesized that with more classes, it is harder to classify the objects, because this demands more discrimination from the partitioning techniques. More categories per variable are hypothesized to lead to more potential variation between objects in the samples, so objects should be better discriminated. A similar argument holds for the number of variables: objects can show more variation if they are measured on more variables, even when the variables are repeated measures, since a probabilistic instead of deterministic mechanism underlies the response patterns scored.

Chaturvedi et al. (2001) find no effect of relative class size on cluster recovery in both LCA and *K*-modes clustering, so no significant main effect for this independent variable is hypothesized in this study. However, Van Buuren and Heiser (1989) note that the *K*-means clustering in GROUPALS tends to partition the data in clusters of roughly equal size, and that if there is prior evidence that this is not the case for the data set under consideration, this technique is not appropriate. Therefore, an interaction effect is hypothesized for the effect of relative class size on the comparison between LCA and GROUPALS, where GROUPALS is expected to have lower cluster recovery than LCA if the relative class size is unbalanced.

2. METHOD

The main purpose of this study is to compare the ability to recover cluster membership of restricted GROUPALS and restricted latent class analysis. Therefore, a simulation study was conducted, where several pseudo-populations were generated. In these pseudo-populations, the cluster each object belonged to was known (the 'true' cluster membership). The populations differed in some aspects such as number of clusters and number of categories of the variables. From each of these populations, 500 random samples ($N = 300$) were drawn, and these data have been analyzed both by restricted GROUPALS and by restricted LCA. This raised the opportunity to compare the 'true' cluster membership recovery of both methods, on the same data sets. Appendix A gives a schematic overview of the steps in the simulation study.

Next, the research design and the dependent variable are specified, and the method used to generate the data is discussed.

2.1 Research design

In the generation of the artificial data, four aspects that could potentially affect the performance of the clustering procedures were systematically varied: the number of classes, the relative size of these classes, the number of variables, and the number of categories each variable has (Table 1).

Table 1. Systematically varied data aspects

data aspect	levels
number of variables	5 and 10
number of classes	3, 4, and 5
number of categories	(3,)[3] 5 and 7
relative class size	balanced and unbalanced

The two levels of relative class size were operationalized as follows. In the balanced class size condition, all classes are of equal size. For the unbalanced class size condition, if the classes are

[3] As discussed in the section on data generation, three categories per variable turned out to be not feasible for populations with more than three classes.

ordered by increasing class size, every class is two times as large as the class preceding it, so the largest class always contains more than 50% of the objects. The number of categories is equal for all variables in the pseudo-population, since the variables were set to have equivalent categories, and of course this is only possible if all variables have the same number of categories.

The full crossing of these data aspects resulted in a 3 x 2 x 3 x 2 design to generate 36 different pseudo-populations. However, as will be further discussed in the section on data generation, it turned out not to be suitable to have more classes than categories of the variables, so in the four- and five classes conditions, there were no data generated with three categories per variable. This means that only 24 cells of the fully crossed design are feasible, each replicated 500 times (the number of samples drawn from each pseudo-population). Four more cells were generated with 3 classes and 3 categories: these are discussed as extensions of the effect for number of categories on cluster recovery.

The data aspects can be viewed as 'between'- factors in the research design. In contrast, all the samples are analyzed twice, so the analyses are 'within' the samples. This means that there is one 'within'-factor: partitioning technique, which has two levels (restricted LCA and restricted GROUPALS).

2.2 Cluster recovery

The main interest in this study lies in the performance of both partitioning techniques in recovering the true cluster membership. It is possible to use an external criterion for cluster recovery in this study, because there is information available on the cluster structure of the pseudo-populations, apart from the clustering process. Several indices for measuring the agreement between two partitions exist. In this study the adjusted Rand index was chosen, of which Saltstone and Stange (1996, p. 169) say that '*it is virtually the consensus of the clustering community that, as an index of cluster recovery for comparing two partitions, Hubert and Arabie's (1985) adjusted Rand index possesses the most desirable properties*'.

The adjusted Rand index is based on two partitions of a set of objects, and assesses for each pair of objects: *(a)* if these objects are in the same class in both partitions, *(b* and *c)* if they are in the same class in one partition, but in different classes in the other partition, or *(d)* if they are in

different classes in both partitions. In this definition, a and d can be interpreted as measures of agreement in the classifications, while b and c are indicative of disagreements. The Rand index (Rand, 1971) is then simply: $(a + d) / (a + b + c + d)$. This index, however, has some unattractive properties. First, it is affected by the presence of unequal cluster size, and second, the expected value of the index of two random partitions is higher than zero, and doesn't take a constant value. Therefore, Hubert and Arabie (1985) proposed to adjust this index for chance, by introducing a null model for randomness in which two partitions are picked at random, subject to having the original number of classes and objects in each. The resulting adjusted Rand index does not suffer from the problems discussed for the (unadjusted) Rand index.

To compute the adjusted Rand index, first a cross tabulation of the two partitions should be constructed, with elements n_{ij} denoting the number of objects that are in class i in the first partition and in class j in the second partition, $n_{\cdot j} = \sum_j n_{ij}$ are the column sums, and $n_{i\cdot} = \sum_i n_{ij}$ are the row sums. In formula, the adjusted Rand index is then:

$$\text{adj.Rand} = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}\right] / \binom{n}{2}} \tag{11}$$

This index takes values between 0 and 1, and it is 0 when the partitions are chosen at random, and it is 1 when the partitions are identical.

For each of the 24 cells of the design, 500 replications of the adjusted Rand value were computed, both for the analyses by restricted GROUPALS and for the analyses by restricted LCA. These values served as the dependent variable. The fact that it is bounded by 0 and 1 may have had consequences for the distribution of the dependent variable, as discussed later in the Results section.

2.3 Data generation

2.3.1 Population and sample size

The size of the samples drawn from the pseudo-populations was chosen to be 300. It is believed that this size is representative for the size of a sample in a real-life study. The size of the pseudo-

population was then determined according to the rule that the sample size should at most be the square root of the size of the population, resulting in a population of at least 90,000 objects. It can be shown that a population of this size is for all practical purposes comparable to an infinite population (De Craen, Commandeur, Frank, & Heiser, in press). From each pseudo-population, 500 random samples were drawn.

2.3.2 *Procedure data generation*

The pseudo-populations were generated according to the latent class model, so the characteristics of this model needed to be specified first. These specifications are the latent class probabilities (or class sizes) and the conditional probabilities for each variable. These conditional probabilities are set equal for all variables, because the variables are assumed to be parallel indicators. Suppose there are two latent classes α and β, and 2 observed variables with both 3 categories. For example, let the latent class probabilities both be .50, and let the conditional probabilities (conditional on class membership) for both variables be as in the following matrix:

$$\begin{array}{c} \\ \textbf{class } \alpha \\ \textbf{class } \beta \end{array} \begin{array}{c} \begin{array}{ccc} \textbf{cat 1} & \textbf{cat 2} & \textbf{cat 3} \end{array} \\ \begin{bmatrix} .10 & .20 & .70 \\ .20 & .70 & .10 \end{bmatrix} \end{array}.$$

Now, it is possible to determine the joint probability of cluster membership and response profile, according to formula (9). For example, the probability of being in class α, scoring category 1 on variable A and also category 1 on variable B = .50 x .10 x .10 = .005.

Data are generated in the following way. First, the latent class probabilities are cumulated over the classes as follows: [.50 1.00]. Then the conditional probabilities are cumulated over the categories:

$$\begin{bmatrix} .10 & .30 & 1.00 \\ .20 & .90 & 1.00 \end{bmatrix}.$$

Now, for each object three random numbers between 0 and 1 are generated. The first number determines the class the object belongs to, and the second and third number determine the categories scored on the first and second variable, respectively. Suppose for some object these random numbers are [.51 .69 .11]. The first number is between .50 and 1.00, so this object is in

class β. To determine the categories scored on the variables, the second row of the matrix with conditional probabilities should be considered, since that is where the probabilities conditional upon being in class β are. The second number .69 is between .20 and .90, so on the first variable this object scores category 2. Similarly, the last number is below .20, so on the second variable this object scores category 1.

This procedure is repeated for all objects in a population, resulting in a population data set consisting of the 'true' class an object is in, and the scores on the variables.

2.3.3 *Determining the conditional probabilities*

In the above it was made clear that to generate the data, the class probabilities and the conditional probabilities should be specified. The class probabilities are part of the research design, and are either balanced (equal for all classes) or unbalanced (each class probability is two times as large as the probability preceding it, if the probabilities are sorted from small to large). The conditional probabilities are not a part of the research design, so these should be determined in some other way.

In a pilot study where the conditional probabilities were determined by random numbers, it appeared that the specific configuration of these conditional probabilities had a very large effect on cluster recovery. This makes sense since two clusters with quite similar conditional probabilities are much harder to recover by cluster analysis than two clusters with quite dissimilar conditional probabilities. For this reason it turned out not to be attractive to determine these conditional probabilities at random. Instead, care should be taken to make the configurations of conditional probabilities comparable in some well defined way for all the populations in the design. Otherwise, if for example an effect on cluster recovery would be found for number of classes, this could be confounded with the effect of the specific conditional probabilities in the respective 3-, 4- and 5-classes case: it could be possible that while there was no real effect of number of classes, it would have seemed like that because the (randomly determined) conditional probabilities in the 3-classes case were more 'suitable' for clustering than those of the 4- and 5-classes case.

As a way to make the conditional probabilities for all populations as similar as possible, it was decided to let the clusters lie on a simplex: the Euclidean distances between the conditional probabilities of all pairs of clusters in a population were set to be equal. This was done in all populations: in all populations the clusters lie on a simplex with vertices $\sqrt{.45} = .6708$. This latter value seemed like a reasonable value in a pilot study, and it is about half of the maximum possible distance between conditional probabilities of clusters, which is $\sqrt{2} = 1.41$. The procedure to derive conditional probabilities such that the clusters lie on a simplex is illustrated now for the case of three clusters and five categories per variable.

Suppose there are three clusters α, β and γ. The matrix with conditional probabilities is as follows:

$$\begin{array}{c} \\ \alpha \\ \beta \\ \gamma \end{array} \begin{array}{c} \begin{array}{ccccc} \textbf{cat 1} & \textbf{cat 2} & \textbf{cat 3} & \textbf{cat 4} & \textbf{cat 5} \end{array} \\ \begin{bmatrix} \pi_{1|\alpha} & \pi_{2|\alpha} & \pi_{3|\alpha} & \pi_{4|\alpha} & \pi_{5|\alpha} \\ \pi_{1|\beta} & \pi_{2|\beta} & \pi_{3|\beta} & \pi_{4|\beta} & \pi_{5|\beta} \\ \pi_{1|\gamma} & \pi_{2|\gamma} & \pi_{3|\gamma} & \pi_{4|\gamma} & \pi_{5|\gamma} \end{bmatrix} \end{array} \begin{array}{c} \\ 1 \\ 1 \\ 1 \end{array} \tag{12}$$

For each cluster, 5 conditional probabilities need to be determined, one for each category. This leads to a total of 15 values that need to be set. However, there are two kinds of restrictions: the probabilities should for each cluster sum to 1, and the (squared) Euclidean distance between all clusters should be the same and equal to .45 ($d^2_{\alpha\beta} = d^2_{\alpha\gamma} = d^2_{\beta\gamma} = .45$).

The requirement that the rows of (10) should sum to 1 leads to a loss of 3 degrees of freedom, and the requirement that the squared distances between all pairs of clusters are equal and fixed to the value .45, leads to another loss of 3 degrees of freedom. This leaves 9 degrees of freedom, so the determination of the values in matrix (12) starts with setting 9 cells to arbitrary values. Since these values represent probabilities, they should be between 0 and 1. In the following matrix (13), a, b and c represent these fixed values, while u, v, w, x, y and z (bold) represent values that need to be determined.

$$\begin{array}{c} \\ \alpha \\ \beta \\ \gamma \end{array} \begin{array}{c} \begin{array}{ccccc} \textbf{cat 1} & \textbf{cat 2} & \textbf{cat 3} & \textbf{cat 4} & \textbf{cat 5} \end{array} \\ \begin{bmatrix} a & b & c & \boldsymbol{x} & \boldsymbol{u} \\ \boldsymbol{v} & a & b & c & \boldsymbol{y} \\ \boldsymbol{z} & \boldsymbol{w} & a & b & c \end{bmatrix} \end{array} \begin{array}{c} \\ 1 \\ 1 \\ 1 \end{array} \tag{13}$$

Note that some rather arbitrary choices have been made: firstly the decision to let the 9 values that were set in advance consist of only three different values a, b and c, and second the positions of these fixed values in the rows of (11). Since the purpose was not to solve the general problem of defining conditional problems in such a way that the clusters lie on a simplex for all situations, but instead to come up with only one solution for this particular situation, it is argued that these arbitrary decisions are not a problem.

Now, it is possible to define a set of equations for the Euclidean distances between the three pairs of clusters. First note that u, v and w are redundant since they can be written as functions of the other parameters in the following way: $u = 1 - a - b - c - x$, $v = 1 - a - b - c - y$ and $w = 1 - a - b - c - z$. So, only x, y and z are variables that need to be solved for. The set of equations is as follows:

$$\begin{cases} d^2_{\alpha\beta} = (1-2a-b-c-y)^2 + (b-a)^2 + (c-b)^2 + (x-c)^2 + (1-a-b-c-x-y)^2 = .45 \\ d^2_{\alpha\gamma} = (b-y)^2 + (1-a-2b-c-z)^2 + (c-a)^2 + (x-b)^2 + (1-a-b-2c-x)^2 = .45 \\ d^2_{\beta\gamma} = (1-a-b-c-y-z)^2 + (1-2a-b-c-z)^2 + (b-a)^2 + (b-x)^2 + (y-c)^2 = .45 \end{cases} \quad (14)$$

This set of three quadratic equations with three variables is not solved easily algebraically by hand, and therefore it was inserted to Maple 9.5, a comprehensive environment for mathematical applications. Maple solves (12) for x, y and z as functions of a, b and c. In case the values $a = .050$, $b = .100$ and $c = .075$ are chosen, $x = .297$, $y = .173$ and $z = .200$. Matrix (13) filled in then becomes:

	cat 1	cat 2	cat 3	cat 4	cat 5	
α	.050	.100	.075	.297	.478	1
β	.602	.050	.100	.075	.173	1
γ	.200	.575	.050	.100	.075	1

These steps were taken for all combinations of number of clusters with number of categories, resulting in the conditional probabilities necessary to generate the pseudo-populations (the second step in the schematic overview in appendix A). The exact conditional probabilities used in the data generation are given in appendix B. In section 4.3.2, potential effects of the specific configuration of the conditional probabilities are explored.

2.4 Analyses

Both in the analyses by restricted GROUPALS and 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 number of clusters minus 1, so that as much information as possible from the data was retained.

2.4.1 Restricted GROUPALS

The algorithm for GROUPALS with equality restrictions on the category quantifications such as described in section 1.4.2 was programmed in MATLAB (Student Version 12). However, the *K*-means algorithm, which is part of the GROUPALS algorithm, is much influenced by local optima. Therefore also GROUPALS (and restricted GROUPALS) is much influenced by local optima. Although Van Buuren and Heiser (1989) noted that the several locally optimal solutions occurring in a test run did not differ to a great extent from the globally optimal solution with respect to the quantifications and cluster means, in test runs in this study, it appeared that the cluster membership of the objects did differ to a considerable extent. In a test run with data with 3 equally sized clusters, 5 variables with all 3 categories ($N = 1000$), 500 restricted GROUPALS analyses were carried out. The adjusted Rand index in this case was in most analyses around 0.73, but there were some outliers with a very low fit and a very low adjusted Rand index of even 0.25.

Although in practice the main interest usually lies in the quantifications and cluster positions instead of in the cluster membership of all the objects, and local optimal solutions may not differ much on these aspects, it is obvious that in this study where the interest does lie in the cluster membership recovery, it is no option to be satisfied with locally optimal solutions. Therefore all the restricted GROUPALS analyses were performed 200 times with different random starting partitions (G_c^0), and the solution with the best fit (the least loss) was chosen as the optimal solution. The agreement of the cluster membership from this solution with the 'true' cluster memberships was computed by means of the adjusted Rand index.

In practice, the true cluster membership of the objects is not known, so the only criterion to choose a 'best' solution is the fit. It is worth mentioning that there is no guarantee that the solution with the best fit is the solution with the best cluster recovery. Moreover, from some pilot studies it appeared that in some cases the fit of the solution need not even to be positively related to the cluster recovery. This was the case when the cluster characteristics in the form of the conditional probabilities were very similar for some clusters. This raised another reason to let the clusters lie on a simplex, where all the clusters are equally dissimilar. In that case, the fit of the solution turned out to be positively related to the cluster recovery, so giving a foundation to pick the solution with the best fit to compute the cluster recovery. In section 4.3.3 the relation between the fit of the solution and the cluster recovery is explored further for one specific cell of the present simulation study.

2.4.2 Restricted LCA

The restricted latent class analyses were carried out with ℓEM (Vermunt, 1997), a general program for the analysis of categorical data. In this program it is possible to do latent class analyses, and to restrict the estimated conditional probabilities to be equal for all variables. In the resulting solution, no attention was paid to aspects such as fit and estimated parameters, since these were not of interest in this study. Only the posterior probability for class membership for the objects was saved, and objects were assigned to the class for which they had the highest posterior probability (modal assignment).

The log-likelihood function that is optimized in restricted LCA, also suffers from local minima. Therefore, all the restricted latent class analyses were carried out twice with different starting values. In that procedure, most of the local optima were captured. Cases with still a very low cluster recovery indicated a local minimum of the log-likelihood, and for these samples several more restricted latent class analyses were carried out, until the algorithm appeared not to converge to a local optimum anymore.

This procedure could not guarantee that all restricted LCAs have reached their global optimum, but at least all severe local optima were captured, and probably most of the less severe

local optima too. A more extensive procedure where all the LCAs were carried out several times, turned out not to be practically feasible.

2.4.3 Analyses on cluster recovery

The obtained indices for cluster recovery were analyzed with a five-way (one within, four between) repeated measures analysis of variance (ANOVA). Furthermore, the data aspects were extended to more levels for some specific, well-chosen situations. For example, the number of categories was extended to include three categories per variable, but only in the case of three classes (see earlier discussion on data generation why this was only possible in case of three classes).

The tests for the main effect of the repeated measures variable and of the data aspects provided the answers to the first 5 research questions, the next four research questions were explored by the first-order interactions of the 'between'-factors with the 'within'-factor partitioning method. The final research question explores whether there were any other (higher-order) interactions that are statistically significant and meaningful.

3. RESULTS

The results are presented as follows. First, the data were screened for possible violations of the assumptions of repeated measures ANOVA. Next, the results of the repeated measures ANOVA on the fully crossed 24 - cell design are discussed. In this ANOVA, there were two levels of number of variables (5 and 10), three levels of number of classes (3, 4 and 5), two levels of number of categories per variable (5 and 7) and two levels of relative class size (balanced and unbalanced). Finally, on the basis of found meaningful effects, for specific situations some systematically varied data aspects were extended to more levels. This gave an opportunity to study some effects more comprehensively, and gives starting points for future research.

3.1 Data screening

The complete simulation study consisted of 24 pseudo-populations or cells (see discussion on research design in section 2.1). Each cell was replicated 500 times, since from each population 500 random samples were drawn. This led to a total of 24 x 500 = 12,000 samples or cases. Each case had a score on two dependent variables: the cluster recovery by restricted GROUPALS and the cluster recovery by restricted LCA, both measured with the adjusted Rand index.

To provide answers to the research questions, the data were analyzed by a repeated measures ANOVA. Two approaches exist: the univariate approach and the multivariate approach also called profile analysis (Stevens, 1990). However, in the present study with two repeated measures, these approaches lead to the same results, so the results for one approach (the univariate) are reported here. There were four assumptions and several practical issues that needed to be checked before carrying out a repeated measures ANOVA (Stevens, 1990; Tabachnick & Fidell, 2001). The assumptions were independence of observations, multivariate normality, homogeneity of covariance matrices and sphericity. In the present study, the samples were randomly drawn from each population and therefore the observations are independent. Furthermore, since the cells were filled with a large number of observations and were of equal size, repeated measures ANOVA should be robust to violations of multivariate normality, and also evaluation of homogeneity of covariance matrices is not necessary (Tabachnick and Fidell, 2001). The final assumption of sphericity, requiring that the variances of the differences of all

pairs of repeated measures are equal (not an assumption for the multivariate approach, only for the univariate approach) is not an issue in the present study with only two repeated measures, since there is only one pair of repeated measures.

Some further practical issues were missing data, outliers, linearity and multicollinearity or singularity (Tabachnick and Fidell, 2001). There were no missing data in the present study, and a check on the relation between the two repeated measures for all the cells separately revealed no clear deviations from linearity, nor was there any sign of multicollinearity or singularity. However, as discussed in section 2.4.2, in almost every cell of the research design, there were some extraordinarily low cluster recoveries for restricted LCA, probably due to convergence to a bad local optimum. Those were low both in reference to the indices by restricted LCA for that cell (univariate outlier), and to the value of the index by restricted GROUPALS for that specific sample (multivariate outlier). Since ANOVA is very sensitive to outliers, the following action was taken: the restricted LCAs were repeated for these cases until the algorithm no longer converged to a local optimum. After this procedure was carried out for all cells, there were no severe outlying cases left, so outliers no longer posed a problem for carrying out a repeated measures ANOVA.

Locally optimal solutions were obtained in 7.4% of the samples, but this appeared not to be equal for all cells of the design. In appendix C an impression of the incidence of local optima is given, separately for each cell. It seems that with data with more underlying classes there was a higher frequency of converging to a local minimum (incidence of .017, .046, and .160 in the 3-, 4-, and 5-classes cells, respectively), which could be caused by the fact that more parameters had to be estimated. However, relative class size (incidence for unbalanced classes .020, for balanced classes .086) also seemed to affect the frequency of occurrence of local minima, although that aspect does not affect the number of parameters to be estimated.

3.2 Results on research questions

Before presenting the main results, some issues on testing of the effects for significance should be discussed first. 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. Indeed, almost all

effects (even the five-way interaction effect) turned out to be significant on the .01-level. It can be said that the design is overly powerful, and some significant effects might be trivial and of no practical importance.

To asses the *magnitude* of the effects (as opposed to the *reliability* of the effects assessed by the *p*-value), one should look at the effect size or strength of association (Tabachnick & Fidell, 2001). In ANOVA, a suitable measure for the strength of association is partial η^2:

$$\text{partial}\,\eta^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \tag{15}$$

Note that the sum over all the effects of the partial η^2s can be greater than 1 (as opposed to the standard η^2), so partial η^2s cannot be interpreted as variance accounted for by the effect. However, it is possible to asses the relative importance of the effects tested for. Also, Cohen (1988) has given some guidelines as to what effects (measured with η^2 instead of partial η^2) should be characterized as 'small' ($\eta^2 = .01$), 'medium' ($\eta^2 = .06$) and 'large' ($\eta^2 = .14$). Applying these conventions to partial η^2, which is larger than η^2, will probably assure that all meaningful effects are discovered and interpreted.

Relevant means and standard deviations for the main effects (discussed in section 3.2.1 and 3.2.2) and first order interaction effects with partitioning technique (discussed in section 3.2.3) are displayed in Table 2. In the following, frequent references to these statistics are made.

Table 2. Means and standard deviations (bracketed) of the adjusted Rand index

partitioning technique	number of variables		number of classes			number of categories		relative class size		total
	5	10	3	4	5	5	7	bal.	un-bal.	
restricted GROUPALS	.631 (.094)	.893 (.053)	.789 (.136)	.786 (.140)	.713 (.163)	.737 (.154)	.788 (.144)	.792 (.126)	.733 (.168)	.763 (.151)
restricted LCA	.720 (.074)	.934 (.029)	.850 (.103)	.840 (.111)	.791 (.138)	.815 (.122)	.839 (.120)	.819 (.122)	.836 (.120)	.827 (.121)
average	.676 (.072)	.914 (.037)	.820 (.116)	.813 (.123)	.752 (.145)	.776 (.133)	.814 (.128)	.805 (.122)	.784 (.140)	.794 (.132)

As mentioned earlier, the dependent variable in this study, the adjusted Rand index, is bounded by 0 and 1. This may have had consequences for the variability of the dependent variable in cells

where the mean adjusted Rand index was close to 1 (or 0, but that was not the case in the present study). In those instances, the variance would probably be lower than in other cells, due to a ceiling effect. This ceiling effect probably also affected the appearance of the effects of the data aspects too, which all seemed to flatten when the mean adjusted Rand index gets closer to 1.

3.2.1 Main effect of repeated measure variable partitioning technique

The first research question (see section 1.7) asks whether restricted LCA or restricted GROUPALS has the highest overall cluster recovery. This is the main effect of the repeated measures variable partitioning technique, and this effect can be found in the first row of Table 3.

Table 3. *Effects of partitioning technique on cluster recovery (main effect and first order interactions: 'within' part of design)*

effect	SS (type III)	df	*F*	*p*	partial η^2
partitioning technique	24.97	1	20690.2	.000	.633
partitioning technique x number of variables	3.18	1	2637.6	.000	.180
partitioning technique x number of classes	.065	2	267.7	.000	.043
partitioning technique x number of categories	1.14	1	940.9	.000	.073
partitioning technique x relative class size	8.68	1	7193.7	.000	.375
error (partitioning technique)	14.46	11976			

The overall effect of partitioning technique on cluster recovery is significant and large (partial η^2 = .633). Means and standard deviations are given in Table 2: the mean adjusted Rand index is substantially higher for restricted LCA as partitioning technique than for restricted GROUPALS. Distributions of the adjusted Rand index are displayed graphically in Figure 1. For both partitioning techniques, the distributions are a bit negatively skewed, but seem to have quite equal variances. It is obvious that restricted LCA reached overall higher cluster recovery than restricted GROUPALS. There still seem to be some samples for restricted LCA with quite low cluster recovery in reference to the other samples, but those were probably caused by the skewness of the distribution of cluster recovery due to a ceiling effect.

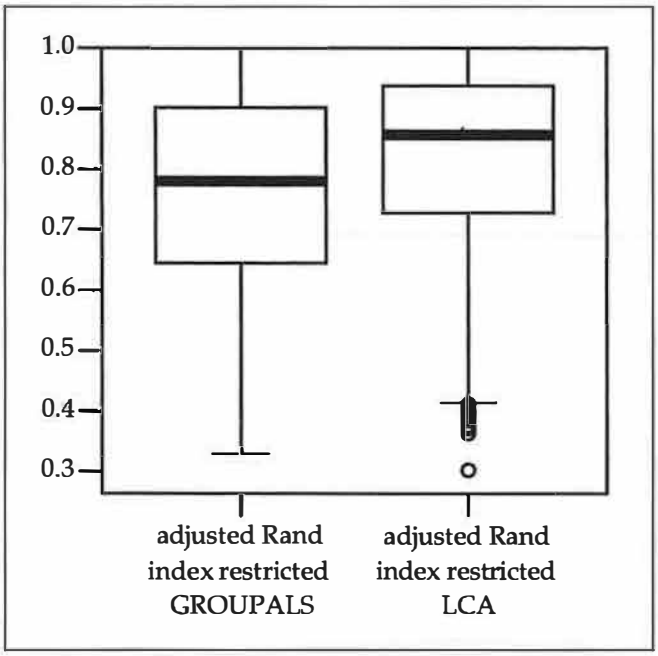


Figure 1. *Boxplots overall adjusted Rand index restricted GROUPALS and restricted LCA*

3.2.2 *Main effects of data aspects*

Research questions 2 to 5 ask for the effects of the systematically varied data aspects on the overall (average) cluster recovery. These aspects were number of variables, number of classes, number of categories per variable and relative class size, and were analyzed by the 'between' part of the repeated measures ANOVA. The tests of these effects are displayed in Table 4.

Table 4. Main effects of data aspects on average cluster recovery ('between' part of design)

effect	SS (type III)	df	*F*	*p*	partial η^2
number of variables	339.42	1	109188.0	.000	.901
number of classes	22.07	2	3459.5	.000	.372
number of categories	8.51	1	2736.2	.000	.186
relative class size	2.61	1	842.6	.000	.066
error	37.23	11976			

All main effects of the data aspects are significant and have a medium to large effect size. The number of variables in the pseudo-population has the largest effect on cluster recovery (partial

η^2 = .901). Looking at the means for the two levels of number of variables, it is obvious that increasing the number of variables increases the overall cluster recovery substantially, as hypothesized: the overall mean adjusted Rand index increases from .676 to .914. This is also displayed graphically in Figure 2, in the first plot.

The effect of number of classes is also significant and large (partial η^2 = .372, second largest effect), and as can be seen in Table 4 and in the second plot in Figure 2, overall cluster recovery decreases when there are more classes underlying the pseudo-population, as was hypothesized. However, there seems to be more decrease on cluster recovery in going from 4 to 5 clusters than from 3 to 4 clusters. This effect is later separated for the partitioning techniques in the discussion on the interaction effect of partitioning technique with number of classes.

Increase of the number of categories per variable leads to a higher overall cluster recovery (partial η^2 = .186, effect displayed in third plot Figure 2), which is also a large effect. This effect is also in the hypothesized direction.

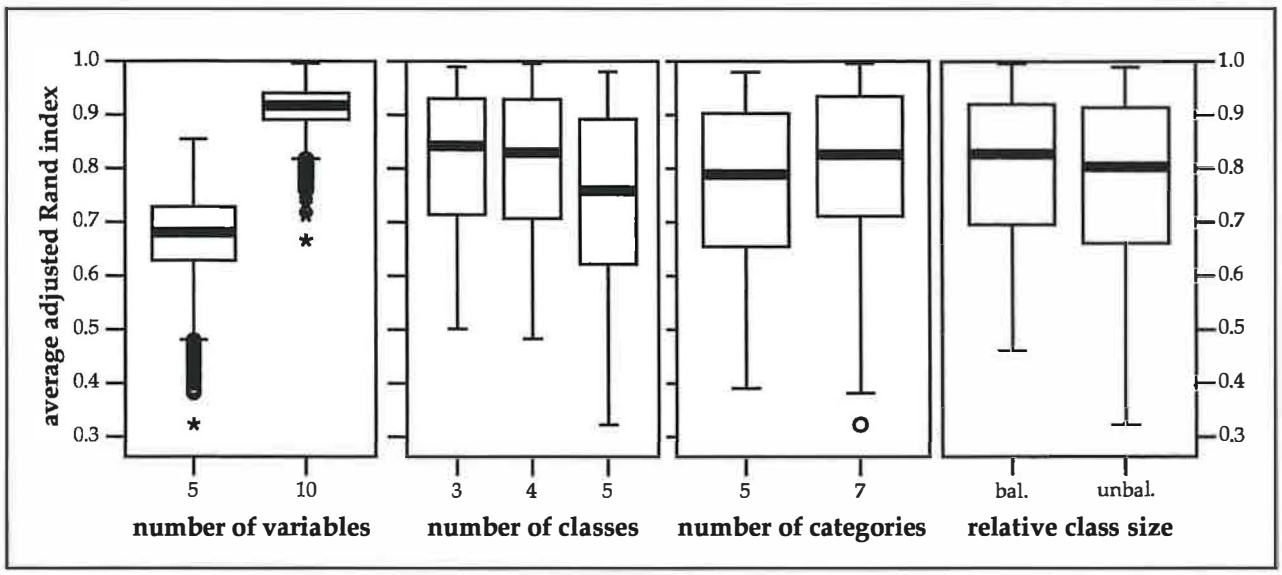


Figure 2. Boxplots for number of variables, number of clusters, number of categories and relative cluster size on the average cluster recovery

Finally, the relative class size has a significant but medium effect on overall cluster recovery (partial η^2 = .066). From Table 2 and the last plot in Figure 2, it becomes clear that overall cluster recovery is larger if the clusters are balanced than if they are unbalanced in size. It was

hypothesized that this would be predominantly the case for restricted GROUPALS, since the incorporated *K*-means algorithm is known to find clusters of roughly equal size. Whether this is the case is explored in the discussion on the interaction effect of partitioning technique with relative class size.

3.2.3 Interaction effects of data aspects with partitioning technique

As followed from section 3.2.1, restricted GROUPALS resulted in lower cluster recovery than restricted LCA. Furthermore, from section 3.2.2 it appeared that the number of variables and the number of categories were positively related to cluster recovery, while the number of classes was negatively related to cluster recovery. Finally, unbalanced class sizes resulted in slightly lower cluster recovery than balanced class sizes.

This section is about possible interaction effects between the data aspects and the partitioning technique (research questions 6 to 9). Statistics of these effects are displayed in Table 3. Again, all effects are significant, but this time range in size from small/medium to large.

There is a large interaction effect between the partitioning technique and number of variables (partial η^2 = .180). Looking at the means in Table 2, displayed graphically in the first plot in Figure 3, it appears that although restricted LCA obtains a higher cluster recovery than restricted GROUPALS both with 5 and with 10 variables, this difference is less marked when the data consisted of 10 than of 5 variables. Worded differently, restricted GROUPALS seems to benefit more from an increase in number of variables than restricted LCA does. In section 3.3.1 it is explored whether this interaction effect is similar when data sets with 15 variables are also included.

The interaction effect between the partitioning technique and the number of classes is small to medium (partial η^2 = .043). The second graph in Figure 3 reveals that the earlier discussed effect that an increase from 3 to 4 classes resulted in less decrease in cluster recovery than an increase from 4 to 5 classes did, can predominantly be ascribed to restricted GROUPALS, and less so to restricted LCA. For restricted GROUPALS, cluster recovery is nearly equal for data sets with 3 classes as for data sets with 4 classes with mean adjusted Rand indices of .789 and .786 respectively, but it decreases to .713 in the 5 classes case. Cluster recovery for restricted LCA

follows a similar pattern but has a more evenly spread decrease in cluster recovery from .850 to .840 to .791 for data sets with 3, 4 and 5 classes, respectively. In section 3.3.2 inclusion of pseudo-populations with 2 underlying classes is studied.

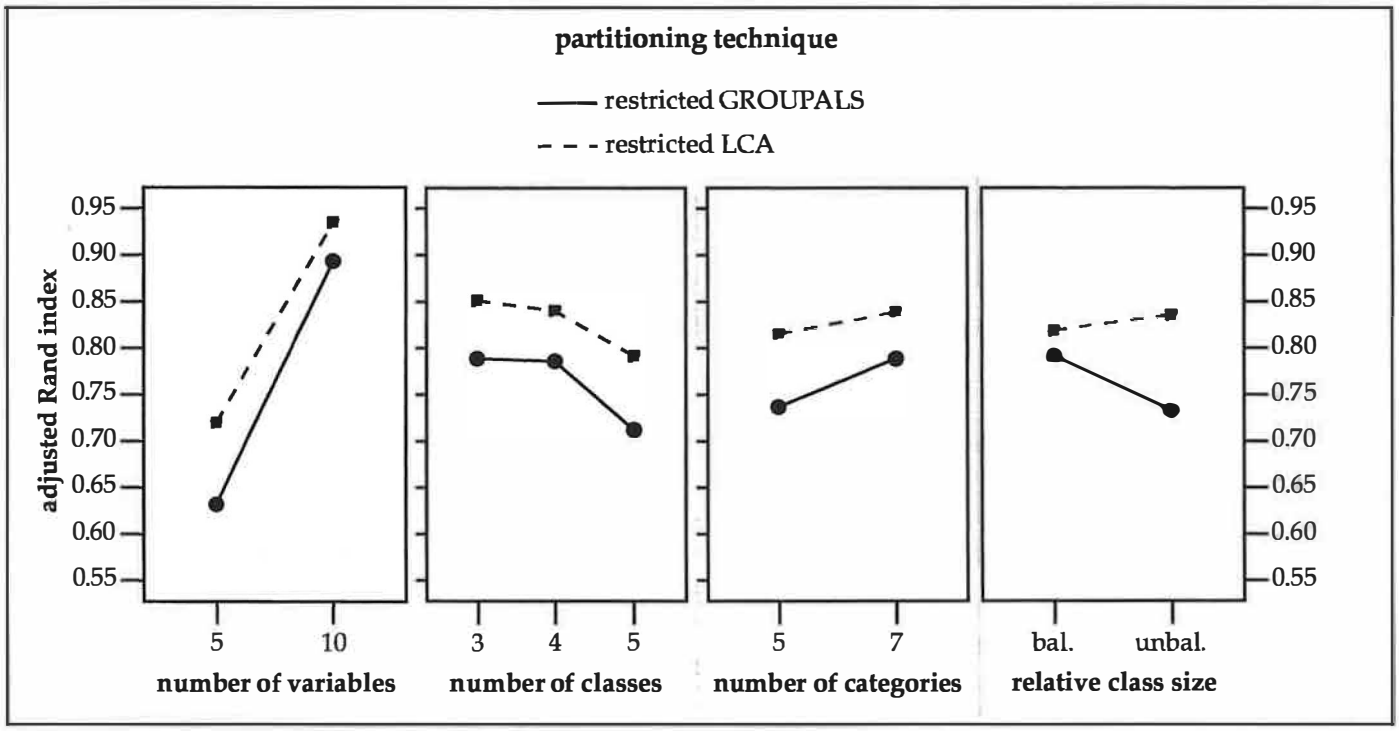


Figure 3. *Mean adjusted Rand index of the interaction effects of partitioning technique with data aspects.*

Partitioning technique is also significantly interacting with number of categories and this effect is of medium size (partial $\eta^2 = .073$). The third graph in Figure 3 displays this effect graphically. As discussed earlier, an increase in number of categories per variable leads to an increase in cluster recovery, but this is even more so for restricted GROUPALS than for restricted LCA. Worded differently, restricted GROUPALS seems to benefit more from an increase in number of categories than restricted LCA does. In section 3.3.3, the number of categories per variable is extended with the inclusion of three categories, so it can be explored whether this effect is similar for 3, 5 and 7 categories per variable.

Finally, there is a large interaction effect between partitioning technique and relative class size (partial $\eta^2 = .375$, the largest effect size of all of the interactions). As can be seen in the last plot in Figure 3, the direction of this effect is partly as hypothesized: restricted GROUPALS has a

substantially higher cluster recovery if the classes are balanced of size than if they are unbalanced, with respective mean adjusted Rand indices of .792 and .733. This is most probably due to the *K*-means clustering algorithm incorporated in restricted GROUPALS. However, cluster recovery by restricted LCA is higher for the unbalanced than for the balanced class size data sets, with mean adjusted Rand indices .819 and .836 respectively, an effect which is hard to explain. In section 3.3.4, 'unbalancedness' is operationalized in several alternative ways, to shed a more detailed and comprehensive light on this phenomenon.

3.2.4 Further relevant effects

Since all but four of effects of the full factorial model were significant at the .01-level, it was decided to only report those effects that had an effect size of at least medium size (partial η^2 = .06). There were two such effects: the interaction between the number of variables and the number of classes on overall cluster recovery, and the three-way interaction between partitioning technique, number of variables and the relative class size.

The interaction effect between number of variables and number of classes is significant ($F(2, 11976) = 585.1$, $p = .000$) and of medium size (partial $\eta^2 = .085$). It is graphically displayed in Figure 4. The decrease in cluster recovery when the number of classes is larger, is more pronounced if the data consist of 5 variables than when they consist of 10 variables.

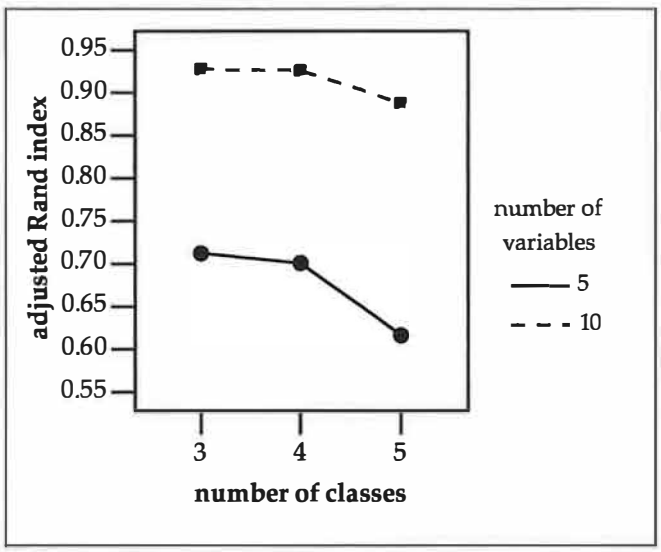


Figure 4. *Mean adjusted Rand index for interaction effect of number of classes with number of variables*

The three-way interaction effect between partitioning technique, number of variables and relative class size is significant ($F(1, 11976) = 2226.6$, $p = .000$) and large (partial $\eta^2 = .157$). Two separate plots were created in Figure 5 to visualize this three-way interaction effect.

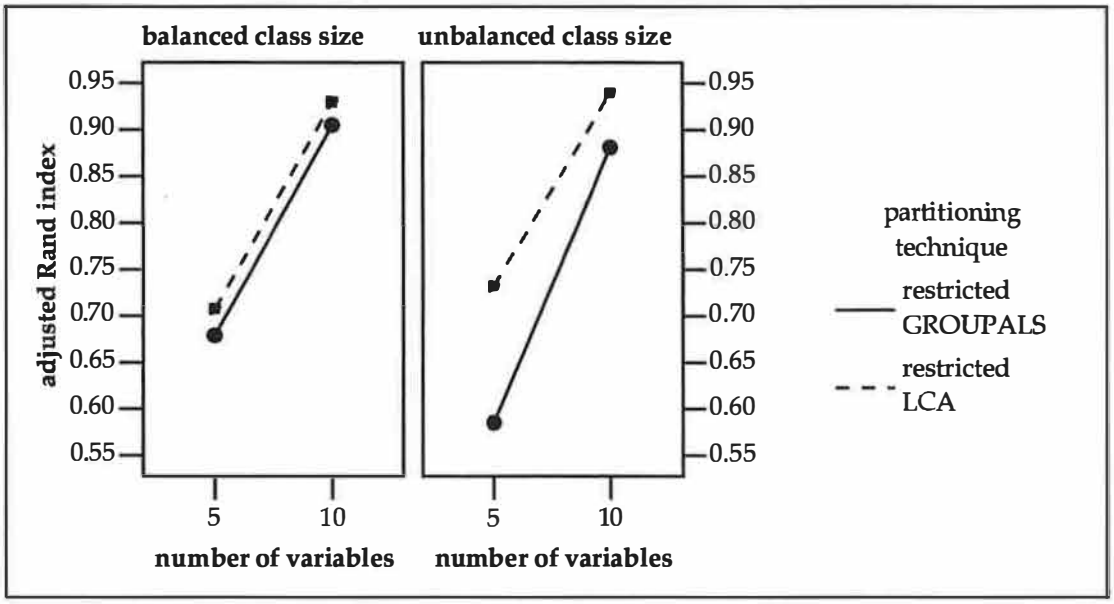


Figure 5. *Mean adjusted Rand index for three-way interaction effect between number of variables, relative class size and partitioning technique.*

As can be seen, the positive effect on cluster recovery of increasing the number of variables from 5 to 10 is approximately of the same magnitude for restricted GROUPALS as it is for restricted LCA, if the cluster sizes are balanced (parallel lines in left plot Figure 5). However, if the cluster sizes are unbalanced, increasing the number of variables from 5 to 10 has a much larger positive effect on cluster recovery for restricted GROUPALS than for restricted LCA (steeper line for restricted GROUPALS than for restricted LCA in right plot Figure 5). However, some notes should be made: firstly, the cluster recovery is at all times larger for restricted LCA than for restricted GROUPALS, and secondly, for the unbalanced class size condition, restricted GROUPALS 'starts' with a quite low cluster recovery for 5 variables, and increases to a more acceptable level for 10 variables.

So, the earlier described effect that restricted GROUPALS benefits more from an increase in number of variables than restricted LCA does, should be ascribed and restricted to the case were the cluster sizes are unbalanced. Worded differently, restricted GROUPALS is not performing

very well in recovering the clusters with 5 variables and unbalanced cluster size, but this is less so if there are 10 variables in the data to be analyzed. If the classes are balanced in size, restricted GROUPALS performs slightly but substantially (effect size = .418) worse than restricted LCA, but this time that difference is independent of the number of variables.

3.3 Relevant extensions of substantial effects to more levels of data aspects

As mentioned in the introduction of the results section and in the discussion of results on the research questions, for some well-defined situations more levels of the systematically varied data aspects were included, to extend the results of the found substantial effects beyond the original fully crossed design. This was done for the number of variables (extension to 15 variables in the data sets), the number of classes (data sets with 2 underlying classes were included), the number of categories (data sets with 3 categories per variable), and for relative class size (unbalancedness was operationalized in several different ways).

3.3.1 Extension of the number of variables

The effect of the number of variables on cluster recovery was very large in the full factorial model: increasing the number of variables from 5 to 10 resulted in a large increase in cluster recovery, and restricted GROUPALS benefited more from more variables than restricted LCA did when the classes were unbalanced in size. When the classes were balanced in size however, this increase was of the same magnitude for restricted GROUPALS as for restricted LCA. To explore whether this effect continues in the same direction when the data contain still more variables, also data with 15 variables were included. The relative class size was varied, since that aspect showed a substantial (three-way) interaction effect with the number of variables and the partitioning technique. The number of categories was held constant at 5, and the number of classes at 3.

First, the main effect of the number of variables with three levels (5, 10 and 15) is significant ($F(2, 2994) = 22713.6$, $p = 0.000$) and very large (partial $\eta^2 = .938$). Increasing the number of variables from 5 to 10 to 15 leads to an overall increase in mean adjusted Rand index from .694 to .918 to .976, respectively. So, the positive effect of number of variables prolongs for more

variables and is also probably affected by a ceiling effect: the adjusted Rand index cannot be higher than 1.00. Now, it is interesting whether this effect is the same for both partitioning techniques (interaction of partitioning technique with number of variables) and whether that interaction effect is the same for balanced and unbalanced class sizes (three-way interaction effect), because those effects were substantial in the original full factorial model.

The interaction effect of partitioning technique with number of variables is significant ($F(2, 2994) = 1004.2$, $p = 0.000$) and large (partial $\eta^2 = .401$), and is shown in the left plot in Figure 6. Again, restricted GROUPALS benefits more from an increase in number of variables than restricted LCA, and this is more so for increasing the number of variables from 5 to 10, than from 10 to 15. However, the second and third plot of Figure 6 show the three-way interaction effect of partitioning technique, number of variables and relative class size. As before, this is a significant ($F(2, 2994) = 316.7$, $p = 0.000$) and large effect (partial $\eta^2 = .175$). Restricted GROUPALS benefits more from an increase in number of variables, but only when the classes are unbalanced in size. When the classes are balanced, increasing the number of variables has (almost) the same effect on restricted GROUPALS as on restricted LCA. This interpretation is similar to that when only 5 or 10 variables were included, so this effect extends in the same direction for 15 variables, although reaching a ceiling of cluster recovery there.

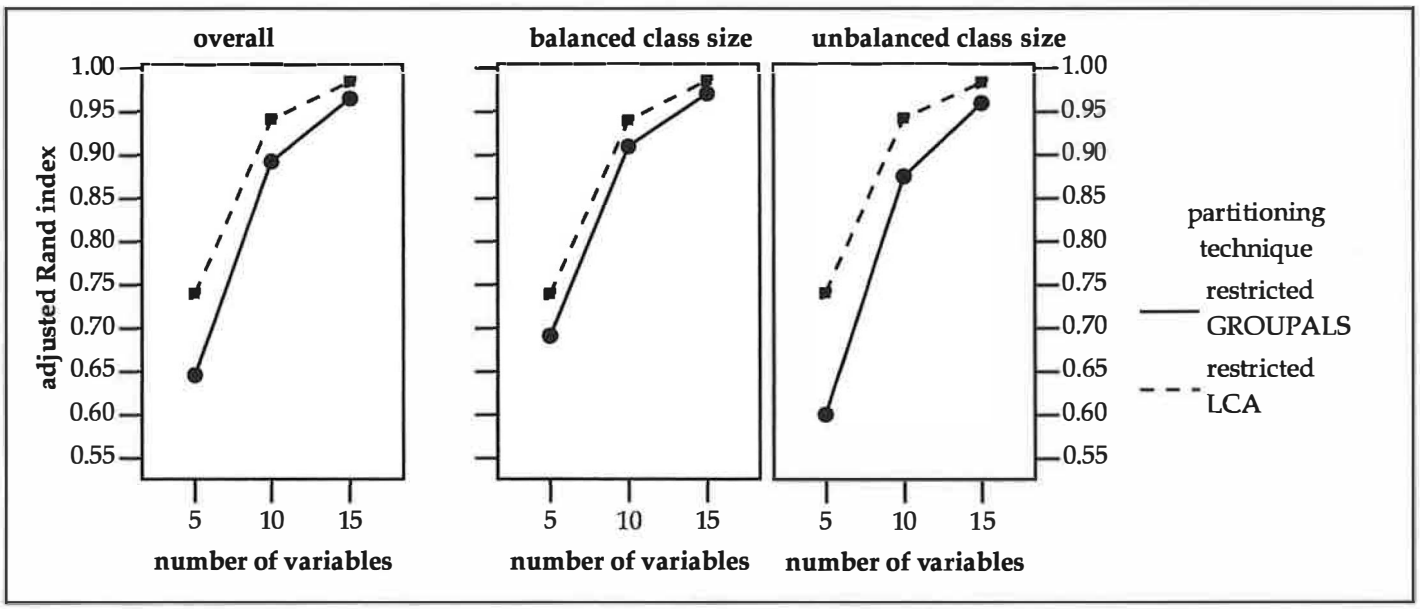


Figure 6. *Mean adjusted Rand index for number of variables (5, 10 or 15) interacting with partitioning technique and in three-way interaction with relative class size; data with 3 classes and 5 categories per variable*

3.3.2 Extension of the number of classes

In the earlier discussed full factorial model, a substantial negative effect of the number of classes on cluster recovery was found. It appeared that an increase from 3 to 4 classes resulted in less decrease in cluster recovery than an increase from 4 to 5 classes did, and that this effect predominantly could be ascribed to restricted GROUPALS, and less so to restricted LCA (second plot in Figure 2). Now, it is explored whether this effect prolongs for 2 classes. It was decided to keep the number of categories per variable constant at 5 and to let the classes be of balanced size, since these two data aspects did not show an interaction effect with the number of classes in the earlier analysis. The number of variables was varied (5 or 10), because that aspect did interact with the number of classes (see section 3.2.4). The interaction effect of the number of variables with partitioning technique turned out to be small (partial $\eta^2 = .021$). This is as expected, since it was observed earlier (in section 3.2.1 and 3.3.1) that with balanced classes, increasing the number of variables had the same effect on cluster recovery for both partitioning techniques.

The main effect of number of classes is significant ($F(3, 3992) = 1930.1$, $p = .000$) and larger than without the inclusion of 2 classes (partial $\eta^2 = .592$). The mean cluster recovery decreases from .871, .819, .793 to .750 for 2, 3 4 and 5 classes, respectively. More interesting is the interaction of the number of classes with the partitioning technique, which is a significant ($F(3, 3992) = 367.0$, $p = .000$) and medium effect (partial $\eta^2 = .059$), displayed in Figure 7.

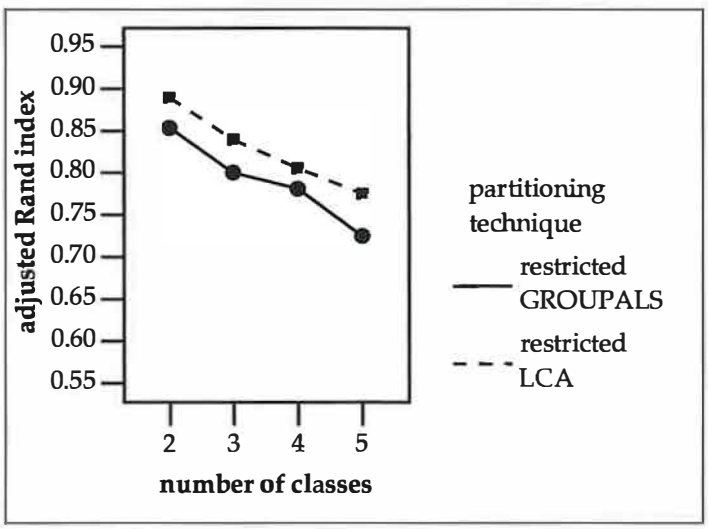


***Figure 7**. Mean adjusted Rand index for the interaction of number of classes with partitioning technique; data with 5 categories and balanced class sizes and number of variables varied*

It seems that for restricted LCA, more classes result in lower cluster recovery, but this decrease is quite evenly distributed. However, the decrease in cluster recovery for restricted GROUPALS is less uniform, and shows the largest decrease when going from 4 to 5 classes. Further research should assess what the cluster recovery is with 6 and more classes.

Furthermore, the number of classes is substantially interacting with the number of variables ($F(3, 3992) = 358.2$, $p = .000$, partial $\eta^2 = .212$), and this effect is in the same direction as it was with only 3, 4, and 5 classes, as can be seen in Figure 8. In data with 10 variables there is less decrease in cluster recovery from an increase in number of classes, than there is in data with 5 variables. Furthermore, the decrease for data with 5 variables is less evenly spread and less uniform.

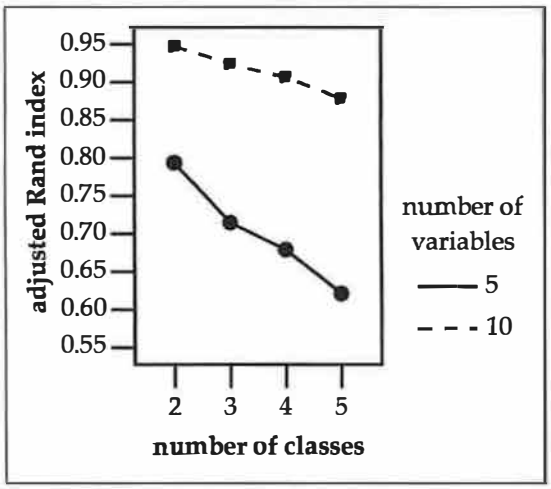


Figure 8. *Mean adjusted Rand index for interaction number of classes with number of variables; data with 5 categories and balanced class sizes*

3.3.3 Extension of the number of categories

Earlier described effects for the number of categories were that cluster recovery increases as the number of categories per variable increased from 5 to 7, and that this is more so for restricted GROUPALS than for restricted LCA, although restricted LCA at all times attains higher cluster recovery (third plot Figure 3). Now, the number of categories is extended to include 3 categories per variable. Under the assumption that the clusters lie on a simplex, this is only possible if the data sets have 3 underlying clusters. Since the number of categories was not interacting substantially with the number of classes, it was decided that it is probably sufficient to only study the effect of having 3 categories per variable for the 3-classes case. What was varied was

the number of variables and the relative class size, although these two aspects were not interacting substantially with the number of categories either, so it would not have been strictly necessary to vary these aspects. So, the factors varied in this subdesign were the number of variables (5 or 10), the number of categories (3, 5, or 7) and the relative class size (balanced or unbalanced.

The main effects of partitioning technique, number of variables and relative class size and their interaction effects were not notably different from those reported in the fully crossed design and are therefore not discussed here. Of interest were the main effects and interaction effects of number of categories, since now also data with 3 categories per variable were included. The main effect of number of categories is significant ($F(2, 5988) = 430.7$, $p = 0.000$) and large (partial $\eta^2 = .677$). The mean cluster recovery increases from .705 for 3 categories to .804 for 5 and to .835 for 7 categories per variable, so the largest gain in cluster recovery is in increasing from 3 to 5 categories.

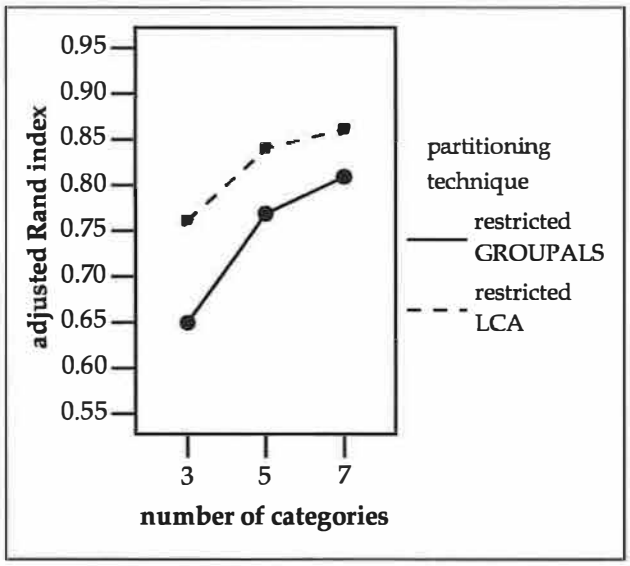


***Figure 9**. Mean adjusted Rand index for interaction of number of categories (3, 5 or 7) with partitioning technique; data with 3 classes, number of variables and relative class size varied*

The interaction effect of number of categories with partitioning technique is also significant ($F(2, 5988) = 906.3$, $p = 0.000$) and large (partial $\eta^2 = .232$), and it is visualized in Figure 9. It seems that

the earlier described trend that restricted GROUPALS benefits more from an increase from 5 to 7 categories is even more so for an increase from 3 to 5 categories.

3.3.4 Extension of relative class size

Relative class size is not a uniquely defined concept, so it can be operationalized in several different ways. Before, it was chosen to study two levels: balanced classes where all classes are of equal size, and unbalanced classes where each class is two times larger than the class preceding it. To get some more insight in the effects of relative class size on cluster recovery, especially on the cluster recovery by restricted GROUPALS where it seems to be a very important factor, unbalancedness was operationalized in several other ways. Instead of the classes being 2 times larger than the class preceding it, unbalanced classes were also generated that were 1.25, 1.50, 1.75 and 3 times larger than the class preceding it. Of course, this is only an extension of the concept unbalancedness in the same line of thought as before, and many more ways can be thought of.

It was decided to keep the number of categories constant at 5 and the number of classes constant at 3, since those two data aspect did not show a substantial interaction effect with relative class size. The number of variables was varied (5 or 10), because that did interact with relative class size in the full factorial model. In total, there were six levels of relative class size: unbalanced x 1.25, x 1.50, x 1.75, x 2 and x 3, and balanced. The main effect of relative class size is significant ($F(5, 5988) = 303.7$, $p = .000$) and large (partial $\eta^2 = .202$), which is much larger than when only two levels of relative class size were included. Overall, the cluster recovery decreases when the classes get more unbalanced in size: the mean adjusted Rand index goes from .819 when the classes are balanced, to .812, .802, .798, .789, and .759 for increasing unbalancedness.

The originally found effect that the relative class size did affect the cluster recovery of restricted LCA slightly in a positive way, while affecting the cluster recovery of restricted GROUPALS substantially in a negative way, is also found now. It is a very large (partial $\eta^2 = .473$) and significant ($F(5, 5988) = 1074.4$, $p = .000$) interaction effect, and it is displayed graphically in the left plot of Figure 10. Indeed, more unbalancedness in class sizes affects cluster recovery by restricted GROUPALS negatively, while having a slight positive effect on cluster

recovery by restricted LCA. It seems that for restricted GROUPALS, cluster recovery decreases linearly when the classes get more unbalanced, which is peculiar, since the increasing levels of unbalancedness are operationalized in a multiplicative way.

Furthermore, this effect might not be the same for the number of variables (5 or 10), and indeed the three-way interaction effect between partitioning technique, relative class size and number of variables is again significant ($F(5, 5988) = 249.0$, $p = .000$) and large (partial $\eta^2 = .172$). This is visualized in the second and third plot in Figure 10. For restricted LCA, both for 5 and for 10 variables the relative class size affects the cluster recovery in a slight positive direction. For restricted GROUPALS however, although the negative effect of more unbalanced class sizes on cluster recovery is for both 5 as for 10 variables quite linear, the effect is larger for 5 variables than for 10. More variables seem to alleviate the negative effect of more unbalanced classes in restricted GROUPALS.

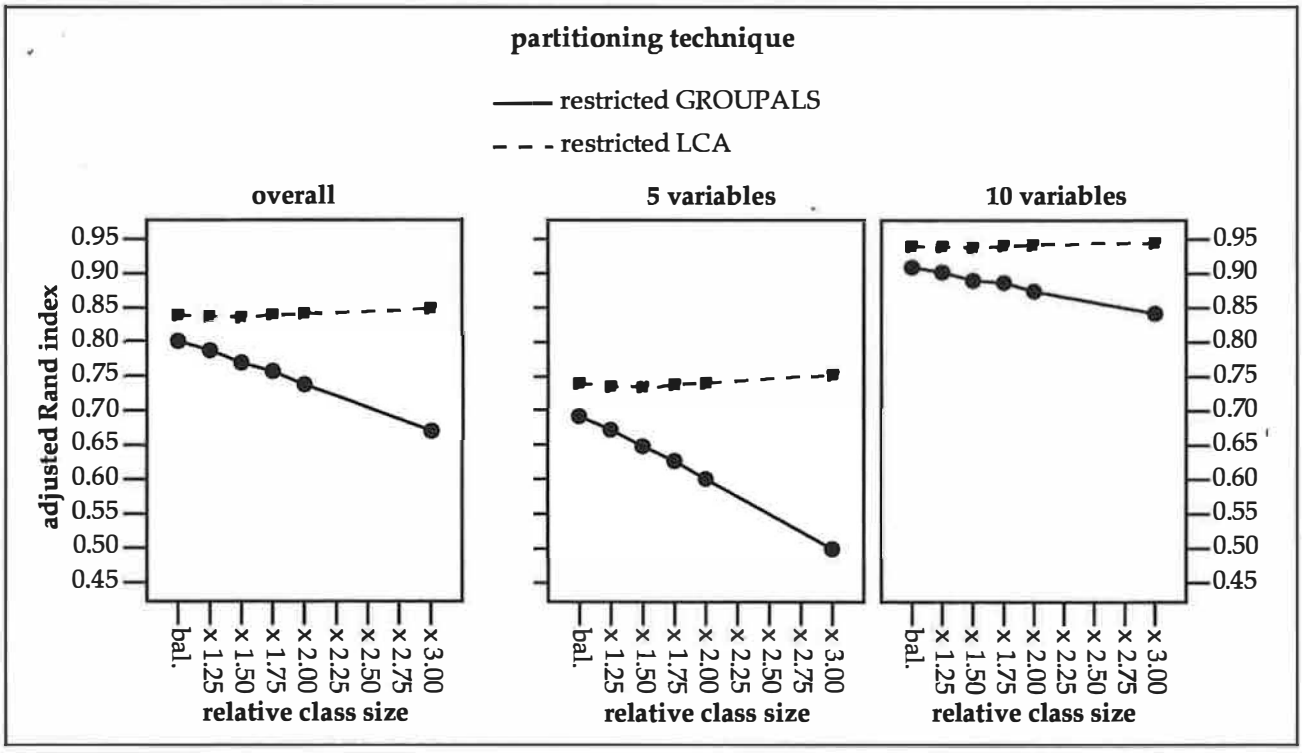


Figure 10. *Mean adjusted Rand index for interaction of partitioning technique with relative class size; data with 3 classes, 5 categories per variable and the number of variables varied*

It was hypothesized that cluster recovery by restricted GROUPALS would be lower when classes underlying the population are unbalanced in size, since the *K*-means algorithm

incorporated in GROUPALS is known to find clusters of roughly equal size. In this simulation study, the 'true' class sizes are known, so it is possible to compare the sizes of the classes estimated by restricted GROUPALS with the 'true' class sizes. In the left plot of Figure 11 this is displayed graphically, the right plot displays the same but for restricted LCA (for one cell of the design: 3 classes underlying the population, data with 5 variables and 5 categories per variable).

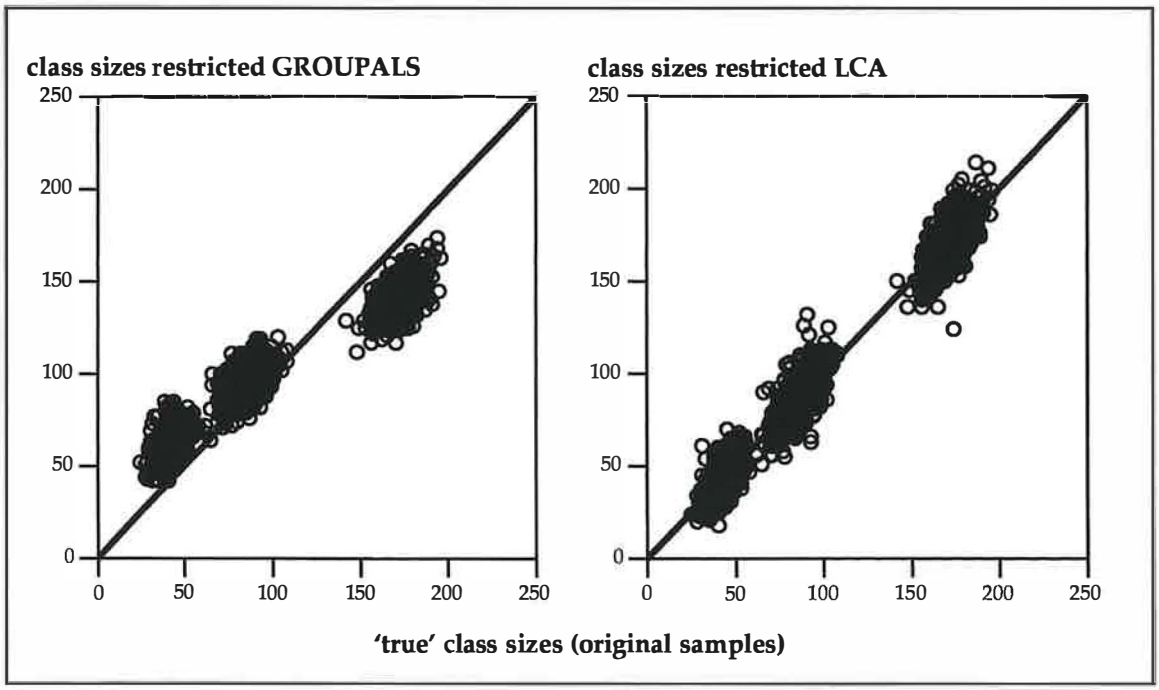


Figure 11. *Relation 'true' class sizes with class sizes found by restricted GROUPALS and restricted LCA*

Indeed, restricted GROUPALS overestimated the sizes of the smallest classes and underestimated the sizes of the largest classes, but this is not as extreme as one might expect. The classes are quite distinct from being of roughly equal size: the largest classes contain about twice as much objects as the smallest classes contain. Restricted LCA seems to find classes that are about the same size as the 'true' classes, although each of the three clouds (per class size) even seems slightly over-positively related with the 'true' class size.

3.4 Further explorations

In previous discussions, several issues came up that deserve some more exploration. Two issues concern the procedures of the simulation study. Firstly, the data generation according to the

latent class model might have given an advantage to restricted LCA over restricted GROUPALS. Therefore, to introduce deviation from the latent class model which LCA estimates, a violation of the local independence assumption was inserted in the data generation. Secondly, in the determination of the conditional probabilities used to generate the data, the decision was made to determine the conditional probabilities such that they lie on a simplex. Several arbitrary choices were made in this process, and now it is explored what the effects of these choices might have been.

Furthermore, two issues concerning GROUPALS are discussed. Firstly the relationship between the fit of a solution and the cluster recovery index is explored. Secondly, it is explored what the cluster recovery might be when instead of the simultaneous scaling and clustering in GROUPALS, a two-step procedure is carried out with first a homogeneity analysis followed by a clustering analysis (*K*-means clustering or model-based clustering).

3.4.1 Data generation revisited: introducing violation of local independence

The artificial data generated according to the latent class model can be viewed as very suitable for LCA, since that is an analysis technique based on the same model according to which the data were generated. This might have given an advantage to restricted LCA over restricted GROUPALS with regard to cluster recovery. To introduce some deviance from the LC model one option is to violate assumptions of the model. Therefore, violation of the assumption of local independence was introduced in the data generation. Local independence means that, conditional upon class membership, the probability to score a category on one variable is independent of the categories scored on the other variables. In the data generation procedure it was therefore allowed to use the same conditional probabilities matrix for every variable (see section 2.3.2).

To introduce local dependence, no longer the same conditional probabilities matrix for all variables sufficed. Conditional on the category scored on the first variable, a different matrix applied for the next variable, and so on for all the variables. These different matrices were constructed by multiplying the original conditional probabilities with random numbers between .7 and 1.7, and normalizing the rows such that for each class, the sum of the conditional

probabilities equaled 1. This procedure was chosen to have some resemblance to the original configurations of the conditional probabilities, so that it would be more justifiable to make comparisons with results from the original data generation.

Preliminary analyses were done for 3 classes of balanced size, 10 variables and 5 or 7 categories per variable. In Table 5 the results are displayed, but note that due to random influences these results cannot be generalized straightforwardly. However, the difference in cluster recovery between restricted GROUPALS and restricted LCA seems to diminish, attenuating the earlier discussed superiority of restricted LCA. Further research should study the effects of different ways of data generation on the relative performance of the two clustering techniques.

Table 5. Comparison of mean adjusted Rand index by original data generation with mean adjusted Rand index by data generation with the assumption of local independence violated

		original data generation			data generation with violation of local independence		
		restricted GROUPALS	restricted LCA	mean difference	restricted GROUPALS	restricted LCA	mean difference
number of categories	5	.909	.938	.029	.917	.936	.020
	7	.935	.949	.014	.916	.920	.003

3.4.2 Determination conditional probabilities revisited

As discussed before in section 2.3.3, the specific configuration of the conditional probabilities used to generate the data had a large effect on cluster recovery. To make comparisons over different cells of the design more justifiable, it was decided to make the conditional probabilities for all cells as comparable as possible in a well-defined way: the classes lie on a simplex with vertices .6708. However, several different configurations meet that requirement, and only one of those for each cell has quite arbitrarily been chosen. It was implicitly assumed that all configurations that met the simplex-requirement would obtain similar cluster recovery results, so it did not matter which one was chosen. Now, it is explored whether this indeed is the case.

For data with 3 underlying classes of balanced sizes, 5 variables and 3 categories per variable, besides the conditional probabilities used in the original data generation, two more configurations that lie on a simplex with vertices .6708 were derived:

version *A* (original)

	cat 1	cat 2	cat 3
α	.100	.635	.265
β	.265	.100	.635
γ	.635	.265	.100

version *B*

	cat 1	cat 2	cat 3
α	.040	.582	.378
β	.378	.040	.582
γ	.582	.378	.040

version *C*

	cat 1	cat 2	cat 3
α	.020	.527	.453
β	.453	.020	.527
γ	.527	.453	.020

In Table 6 statistics on cluster recovery, both by restricted GROUPALS and by restricted LCA, are displayed. A quite striking pattern emerges: the specific configuration does affect cluster recovery, but moreover, it affects cluster recovery by restricted GROUPALS and by restricted LCA in opposite directions. Although in all three versions the distances between all pairs of classes are the same, looking for differences between the versions it seems that in version *A*, there is a quite large difference between the most and the second-most large conditional probability, in version *B* this difference is less and in version *C* this difference is quite small. Potentially, large differences (i.e. one category being the 'dominant' indicator of a class) are beneficial for cluster recovery by restricted GROUPALS, but negatively affect cluster recovery by restricted LCA. The opposite pattern may hold for small differences in most and second-most large conditional probability (i.e. more than one category being the 'dominant' category of a class), and furthermore, cluster recovery by restricted GROUPALS may then show more variability, with a few quite high cluster recoveries. Future research should study this peculiar phenomenon further.

Table 6. *Adjusted Rand index for three different versions of conditional probabilities matrix*

conditional probability matrix	adjusted Rand index								
	restricted LCA				restricted GROUPALS				mean difference
	mean	min	max	SD	mean	min	max	SD	
version A	.641	.427	.784	.061	.514	.360	.660	.049	.127
version B	.784	.628	.884	.043	.391	.281	.821	.095	.393
version C	.832	.714	.925	.039	.278	.234	.711	.075	.554

However, it could have serious implications for the results of the present simulation study. Since apparently the requirement that the classes lie on a simplex does not lead to similar results for all possible configurations, the arbitrary choices made in the determination of the conditional probabilities could have affected cluster recovery by both techniques. Especially the effects for the number of classes and the number of categories might have been contaminated with the

effects of the specific configuration of conditional probabilities used in the generation of the data. Looking at the conditional probabilities used in appendix B however, they seem to be most similar to version *A*, and no striking differences between the several matrices are observed at a first glance, but this does not rule out the possibility of potentially confounded effects. It also raises questions on the generalizability of the results to situations where the conditional probabilities are unlike those in version *A*. However, it is hard to predict the pattern of cluster recoveries on the basis of these preliminary results.

3.4.3 Relation between fit of solution and cluster recovery in restricted GROUPALS

A discussed in section 2.4.1, in restricted GROUPALS the relation between the fit of a solution and the cluster recovery is not straightforward. From some preliminary analyses, when some of the classes underlying the population were quite similar (they could be seen as overlapping), the fit of the solution even seemed to be negatively related with cluster recovery. In the present simulation study, there was not so much overlap between the classes, so it was assumed that the relation between the fit of the solution and the cluster recovery would be positive. Now, for the cell with 3 balanced classes, 5 variables with 5 categories per variable, each sample was analyzed by restricted GROUPALS 200 times, as discussed in section 2.4.1, but instead of only saving the solution with the best fit and computing the cluster recovery of that solution, all solutions were saved and their cluster recoveries were computed. This resulted in 200 pairs of fit and cluster recovery, repeated 500 times (for all the samples). Two aspects were studied: the relation between the fit and the cluster recovery, and the difference between the cluster recovery of the solution with the best fit and the largest cluster recovery for that sample, irrespective of fit.

For all samples, the correlation coefficient between fit and cluster recovery was computed, based on 200 pairs of observations, and indeed the relation in these samples was positive with a mean correlation of .983 and all correlations being larger than .898. Checking the distributions, it appeared that in most samples, there was 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, such as for one sample is displayed in Figure 12. However, the clump with high fit and recovery (for this sample containing 177 of the 200 samples) reveals that there were several analyses that obtained

a high fit (ranging from .668 to .669) but cluster recoveries ranged from .648 to .666. In this clump, there is no significant positive relation between fit and cluster recovery left anymore (correlation coefficient is .095, $p = .207$), so only the few obviously suboptimal solutions were causing the positive relation between fit and cluster recovery.

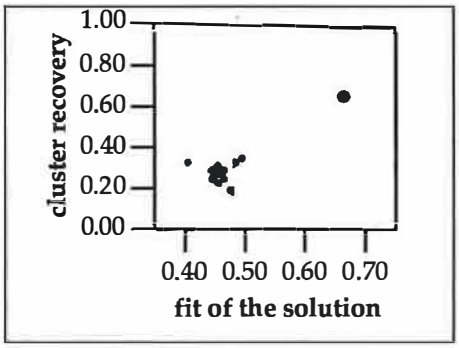


Figure 12. *Relation fit of the solution with cluster recovery of that solution for one sample*

Furthermore, it is theoretically interesting whether there is a difference between the cluster recovery of the solution with the best fit, and the highest cluster recovery irrespective of fit. In practical applications, this usually would not be possible, because 'true' class membership is not known. For this cell with 500 samples, the mean cluster recovery when for each sample the solution with the best fit is chosen as 'best' solution is .691, while when for each sample the solution with highest cluster recovery is chosen, the mean cluster recovery is .708, a mean difference of .017. In Figure 13 this is displayed for all 500 samples.

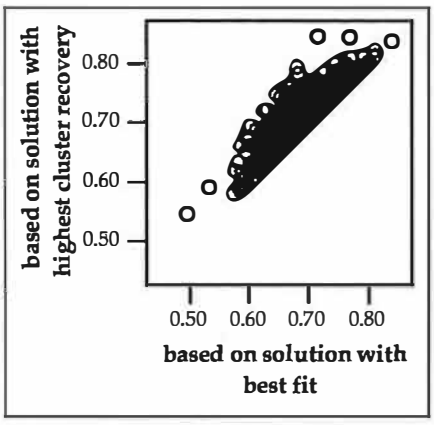


Figure 13. *Cluster recovery of solution based on best fit with cluster recovery of solution based on highest cluster recovery*

The smallest difference is .000 (i.e., the solution with the best fit is also the solution with highest cluster recovery), the largest difference is .128. So, choosing the solution with best fit may be suboptimal with respect to cluster recovery, but in practice, fit of the solution is usually the only criterion available to choose a 'best' solution.

3.4.4 Two-step procedure: multiple correspondence analysis followed by a clustering technique

As discussed in section 1.3, another technique for partitioning categorical data is to apply a two-step procedure: firstly a multiple correspondence analysis to scale the variables, followed by a clustering technique suitable for numerical data. This approach was discredited as unsuitable, because in the first data reduction step, information on the clustering structure might be lost. Simultaneous scaling and clustering was therefore proposed in GROUPALS. Now, it is explored by some preliminary analyses whether these two-step procedures indeed obtain inferior clustering results for the present type of data structures.

In the two-step procedures, the first step of homogeneity analysis had the extra restriction that category quantifications of equivalent categories are equal, similar to the restriction imposed on GROUPALS. As second clustering steps, the *K*-means algorithm and model-based clustering were explored. Model-based clustering postulates a probability model underlying the data: the data are generated by a mixture of probability distributions, conventionally multivariate normal distributions, in which each component represents a different group or cluster (for a discussion of model-based clustering, see for example Fraley & Raftery, 2002). Orientation, volume and shape of the clusters can be specified by posing restrictions on the model, and the *K*-means algorithm can be seen as model-based clustering where the clusters are restricted to be spherical and of equal size (Fraley & Raftery, 2002). Therefore, when the classes are unbalanced in size, a two-step procedure with model-based clustering may outperform a two-step procedure with *K*-means as clustering step and potentially also restricted GROUPALS.

Some preliminary simulations by a two-step analysis with (restricted) homogeneity analysis followed by *K*-means or by model-based clustering were done (MATLAB code for model-based clustering by Martinez & Martinez, 2004). Data ($N = 1000$) consisted of 3 variables with 5 categories and 4 clusters, once balanced in size and once unbalanced in size. The conditional

probabilities used to generate the data were different from those used in the main study, and did not lie exactly on a simplex. All analyses were carried out with 100 different start configurations and the restricted GROUPALS and restricted homogeneity analyses with 3 dimensions. Results are displayed in Table 7.

Table 7. Results preliminary simulations two-step procedures and restricted GROUPALS

clustering technique		mean adjusted Rand index	
		balanced class sizes	**unbalanced class sizes**
restricted GROUPALS		.489	.281
step 1: restricted multiple correspondence analysis	**step 2: *K*-means clustering**	.477	.284
	step 2: model-based clustering	.397	.473

Both for balanced and for unbalanced class sizes, a two-step procedure with *K*-means as clustering step obtained similar recovery results as restricted GROUPALS did, so the simultaneous scaling and clustering may not obtain higher cluster recovery than the two-step procedures.

Furthermore, with unbalanced class sizes, the two-step procedure with model-based clustering as second step obtained substantially higher cluster recovery than with *K*-means as second step and than restricted GROUPALS, supporting the hypothesis that model-based clustering might be better suited if classes are unbalanced in size. When the classes were balanced in size, model-based clustering as second step obtained the lowest cluster recovery. This is contrary to expectance, since *K*-means can be seen as a special (restricted) case of model-based clustering, and it is peculiar that this special case obtained higher cluster recovery than the general, unrestricted case.

In light of these promising results of the two-step procedures of homogeneity analysis followed by a clustering algorithm, discrediting those procedures in advance may not be wise. It would be interesting for future research to explore those sequential approaches in more detail, especially in comparison with results of the simultaneous scaling and clustering in (restricted) GROUPALS.

4. DISCUSSION

In the present study, a new technique to obtain a partitioning of objects measured on several parallel indicators of categorical measurement level was introduced: GROUPALS with equality restrictions on the category quantifications. This technique appeared to obtain reasonable to excellent clustering results, dependent on several factors.

4.1 Conclusions

The purpose of the simulation study was to make a comparison of the cluster recovery of this technique restricted GROUPALS with that of restricted latent class analysis. Several pseudo-populations were generated according to the latent class model, and four data aspects were varied systematically: the number of variables, the number of classes, the number of categories per variable and the relative class size.

Overall, and more importantly in every separate cell of the research design, restricted LCA obtained higher cluster recovery than restricted GROUPALS. However, all the varied data aspects affected cluster recovery in a general way and also differentially for the two partitioning techniques. The most important data aspect appeared to be the number of variables in the data set to be analyzed: increasing the number of variables resulted in a large increase in cluster recovery. This was as hypothesized: more variables result in more possible variation between the objects, so objects can be differentiated better. Moreover, there seemed to be a trend that restricted GROUPALS benefits more from an increase in the number of variables than restricted LCA, but only when the classes underlying the data sets are unbalanced in size. The difference in cluster recovery between restricted GROUPALS and restricted LCA diminished when there were more variables in the data.

Furthermore, it was hypothesized that with data with more underlying classes it is harder to recover the 'true' cluster membership, and this effect was indeed found. Although the presence of more underlying classes resulted in a decrease in cluster recovery for both partitioning techniques, this decrease was for restricted LCA quite linear, while for restricted GROUPALS it was less uniform. A factor influencing the negative effect of the number of classes

on cluster recovery was the number of variables. Increasing the number of variables in the data seemed to alleviate the negative effect of the presence of more classes underlying the data.

Next, it was hypothesized that when the variables in the data have more categories, more variability between the objects is introduced and hence cluster recovery may increase. Indeed, increasing the number of categories per variable led to an increase in cluster recovery, and this was even more so for restricted GROUPALS than for restricted LCA. Furthermore, the effect flattened when cluster recovery reached higher levels with more categories.

Finally, the relative size of the classes was hypothesized to be a factor of influence on cluster recovery by restricted GROUPALS, but not for restricted LCA. Since the *K*-means algorithm incorporated in GROUPALS is known to find classes of roughly equal size, cluster recovery may decrease when the data consist of underlying classes unequal in size. As hypothesized, increasing levels of unbalancedness of the class sizes led to lower levels of cluster recovery for restricted GROUPALS. As mentioned earlier, this negative effect of unbalanced class sizes was less when there were more variables in the data. A peculiar finding was that cluster recovery by restricted LCA seemed to be slightly higher for unbalanced classes than for balanced classes, a finding which is hard to explain and is not found in the literature on latent class analysis.

Summarizing the above, for the data generated according to the way described in section 2.3, restricted LCA obtained higher cluster recovery than restricted GROUPALS. For both techniques, increasing the number of variables and the number of categories per variable positively affected cluster recovery, while the presence of more classes negatively affected cluster recovery. These effects were more pronounced for restricted GROUPALS. Relative class size was predominantly a factor of importance for restricted GROUPALS, where unbalanced classes negatively affected cluster recovery, but it did not affect cluster recovery for restricted LCA much. Finally, increasing the number of variables seemed to alleviate the negative effect of more underlying classes for both techniques and the negative effect of unbalanced class size in restricted GROUPALS.

4.2 Discussion

4.2.1 Limitations

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 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 aspects.

Firstly, one very important factor in the generation of the artificial data is the choice to generate the data according to the latent class model. As discussed in section 3.4.1, these artificial data can be viewed as very suitable for LCA, since that is an analysis technique based on the same model according to which the data were generated. From some preliminary simulations where in the generation of the data the assumption of local independence was violated, the difference in cluster recovery in favor of restricted LCA seemed to diminish, attenuating the discussed superiority of restricted LCA over restricted GROUPALS. This is an important consideration for future research, where it would be very interesting to study the cluster recovery of restricted GROUPALS and restricted LCA with data generated in another way, not based on the LC model, and for practical applications.

A second 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. (2005 in press) did in their simulation study. This was mainly done to make the comparison between different cells of the design more justifiable, especially for the effects of the number of classes and the number of categories. This justifiability can be questioned however, as discussed in section 3.4.2. Moreover, in practical situations all pairs of classes underlying a population do not have to be equally similar or different (which is the case if classes lie on a simplex). So, the question is whether the present results, especially those on number of classes and number of categories per variable, generalize to instances where some classes are quite similar and other classes are quite different from each other. This is also an aspect future research should explore.

Also important in discussing the generalizability of the results are the data aspects and their levels studied. Based on literature such as Chaturvedi et al. (2001), the number of variables, the

number of categories per variable, the number of classes and the relative class size were considered to be potentially affecting cluster recovery. Several other factors such as sample size were not varied. Moreover, only specific levels of these data aspects were included, based on literature, practical considerations and pilot studies. This raises questions about generalizing the results to other levels of these data aspects. Several additional simulations have been performed to study this for some well-defined instances, so going beyond the originally defined research design. Another strong point of the present study was the completely balanced research design where all possible effects of the data aspects could be studied.

One further limitation is that dependent variable of this study was bounded by 0 and 1, so that effects of data aspects on cluster recovery might have been affected by ceiling effects. This may have partially caused the apparent flattening of the effects when cluster recovery reached high values, such as for example was the case for an increase in number of variables. Potentially, this ceiling effect may have been partially responsible for the observations that some effects were more pronounced for restricted GROUPALS than for restricted LCA, because the cluster recovery by restricted GROUPALS 'started' lower, so being less affected by the ceiling. Further research should explore the possibility of transformation of the adjusted Rand index.

One final limitation is that only data without missing values were studied, and that the restricted GROUPALS technique is not yet adjusted to account for missing data, but it should be possible to treat missing data in a way conventional in homogeneity analysis (Gifi, 1990; Van Buuren & Heiser, 1989).

4.2.2 *Issues in restricted GROUPALS*

As mentioned earlier, when some classes were very similar with regard to their conditional probabilities (departure from the simplex class structures, classes showing overlap) the fit of the solution was not positively or even negatively related to cluster recovery. So, solutions that are suboptimal with respect to fit may be optimal with respect to cluster recovery. However, in practical situations cluster recovery usually cannot be assessed, so the only guideline to choose a solution is the fit. Choosing the solution with the best fit may not be appropriate in all instances, and this is an aspect future research should explore. For practical purposes, if there are strong

hypotheses that the underlying classes show a large amount of overlap, restricted GROUPALS might not be the technique of choice. Still, even when the classes were not overlapping much, the solution with the best fit was not at all times the solution with the best cluster recovery.

Furthermore, it was decided to set the dimensionality of the solution in the present study at the maximum number of dimensions, to retain as much information of the original data as possible under assumption of underlying classes. However, it was not studied whether indeed this resulted in higher levels of cluster recovery then when less dimensions were extracted.

Also, the negative effect of cluster recovery of classes that are unbalanced in size on restricted GROUPALS focuses attention on the incorporated *K*-means algorithm. It would be interesting to incorporate some other clustering technique that can handle classes of unequal size and of other shape, such as model-based clustering or fuzzy clustering.

Furthermore, earlier discredited (e.g. Vichi & Kiers, 2001; Van Buuren & Heiser, 1989) two-step procedures with data reduction technique as first step and a clustering technique for numerical data as second step gave promising results. Especially when classes were unbalanced in size, model-based clustering as second step outperformed *K*-means as second step, and also the simultaneous scaling and clustering of restricted GROUPALS. It would be interesting to explore those sequential approaches in more detail.

4.2.3 Issues in restricted LCA

In the present study, the only aspect of LCA that was considered was the obtained classification based on estimated posterior probabilities. Other parameters such as fit, estimated (conditional) probabilities and identification of the model were not studied. Only one dependent variable, cluster recovery, was considered, and although this leads to less complicated results, it also gives a limited perspective of the several aspects that could be studied. One such issue is the frequency of occurrence of local optima in the log-likelihood function, which seemed to be dependent on several factors, such as number of classes and relative class size.

Furthermore, although it was hypothesized that relative class size would not affect cluster recovery by restricted LCA, cluster recovery for unbalanced classes was slightly higher than for balanced classes, although the size of the difference does not imply practical importance.

4.2.4 Recommendations

Finally, some recommendations for practical applications are made. Firstly, a researcher might expect the variables in his or her study to be parallel indicators of some construct, but this is no guarantee that indeed the categories of the variables behave as equivalent. To check this, an unrestricted homogeneity analysis can be performed and it can be explored whether equivalent categories indeed obtain similar quantifications. One might also compare the fit of an unrestricted homogeneity analysis with the fit of a homogeneity analysis where the category quantifications are restricted to be equal. When the fit is not decreasing much by imposing the equality restrictions, this can be seen as support for the presumption that the categories of the variables are equivalent.

When the researcher has established that indeed the variables can be seen as parallel indicators, and he or she wants to derive a partitioning of the objects in different classes underlying the data, several considerations are important. When he or she suspects that the classes are very different in size, restricted LCA is the technique of choice. Also, data with not so many variables (say, 5 or less) and not so many categories per variable (say, less than 5) point in the direction of restricted LCA. However, increasing the number of variables (assuming that the new variables are also similarly behaving parallel indicators of the same construct) diminishes the negative effect of the above data aspects. With, say, 10 or more variables, the differences between restricted LCA and restricted GROUPALS are quite small, with 15 variables they are for practical purposes negligible. So, if one is interested in scaling of the variables in addition to obtaining a clustering of the objects, and there are many variables and the hypothesized classes are not extremely different in size, restricted GROUPALS may be an appropriate technique with very acceptable cluster recovery. Furthermore, if one has reasons to suspect that conditional on class membership, the category probabilities for the different variables are not independent of each other, as for example might be expected with repeated measures variables, restricted GROUPALS may be the technique of choice.

5. 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* (pp. 5 – 64). Singapore: World Scientific Publishing.

Bacher, J. (2000). A Probabilistic Clustering Model for Variables of Mixed Type. *Quality & Quantity, 34*, 223 – 235.

Bock, H.-H. (1996). Probability Models and Hypotheses Testing in Partitioning Cluster Analysis. In P. Arabie, L.J. Hubert, & G. De Soete (Eds.), *Clustering and Classification* (pp. 377 – 454). 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. Second Edition.* Hillsdale NJ: Lawrence Erlbaum Associates.

Collins, L.M., Fidler, P.L., & Wugalter, S.E. (1996). Some Practical Issues Related to the Estimation of Latent Class and Latent Transition Models. In A. von Eye & C.C. Clogg (Eds.), *Categorical Variables in Developmental Research.* London: Academic Press, Inc.

De Craen, S., Commandeur, J.J.F., Frank, L.E., & Heiser,W.J. (2005 in press). Effects of group size and lack of sphericity on the recovery of clusters in K-means cluster analysis. *Multivariate Behavioral Research.*

Everitt, B.S., & Dunn, G. (2001). *Applied Multivariate Data Analysis, Second Edition.* London: Arnold Publishers.

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 Analysis.* Chichester: John Wiley & Sons.

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* (pp.3 – 55). Cambridge: Cambridge University Press.

Gordon, A.D. (1999). *Classification, 2nd Edition.* London: Chapman & Hall.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Hagenaars, J.A., & McCutcheon, A.L. (2002). *Applied Latent Class Analysis*. Cambridge: Cambridge University press.

Hartigan, J.A. (1975). Clustering Algorithms. New York, NY: Wiley.

Heiser,W.J. (1981). *Unfolding analysis of proximity data*. Leiden: Department of Data Theory.

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

Martinez, A.R., & Martinez, W.L. (2004). *Model-Based Clustering Toolbox for MATLAB*. Retrieved at April 24, 2005 from http://www.stat.washington.edu/mclust/

McCutcheon, A.L. (1987). *Latent Class Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-064. Beverly Hills: Sage Pubns.

McCutcheon, A.L. (2002). Basic Concepts and Procedures in Single- and Multiple-Group Latent Class Analysis. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp.56 – 88). Cambridge: Cambridge University Press.

Nishisato, S. (1984). Forced Classification: A Simple Application of a Quantification Method. *Psychometrika, 49*, 25 – 36.

Rand, V.M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association, 66*, 846 – 850.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using Mutivariate Statistics. Fourth Edition*. New York: Allyn and Bacon.

Saltstone, R., & Stange, K. (1996). A Computer Program to Calculate Hubert and Arabie's Adjusted Rand Index. *Journal of Classification, 13*, 169 – 172.

Stevens, J.P. (1990). *Intermediate statistics: A modern approach*. Hillsdale (NJ): Erlbaum.

Van Buuren, S. (1986). *GROUPALS: A Method to Cluster Objects for Variables with Mixed Measurement Levels*. Research Report – 86 – 10. Leiden: Department of Data Theory.

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.

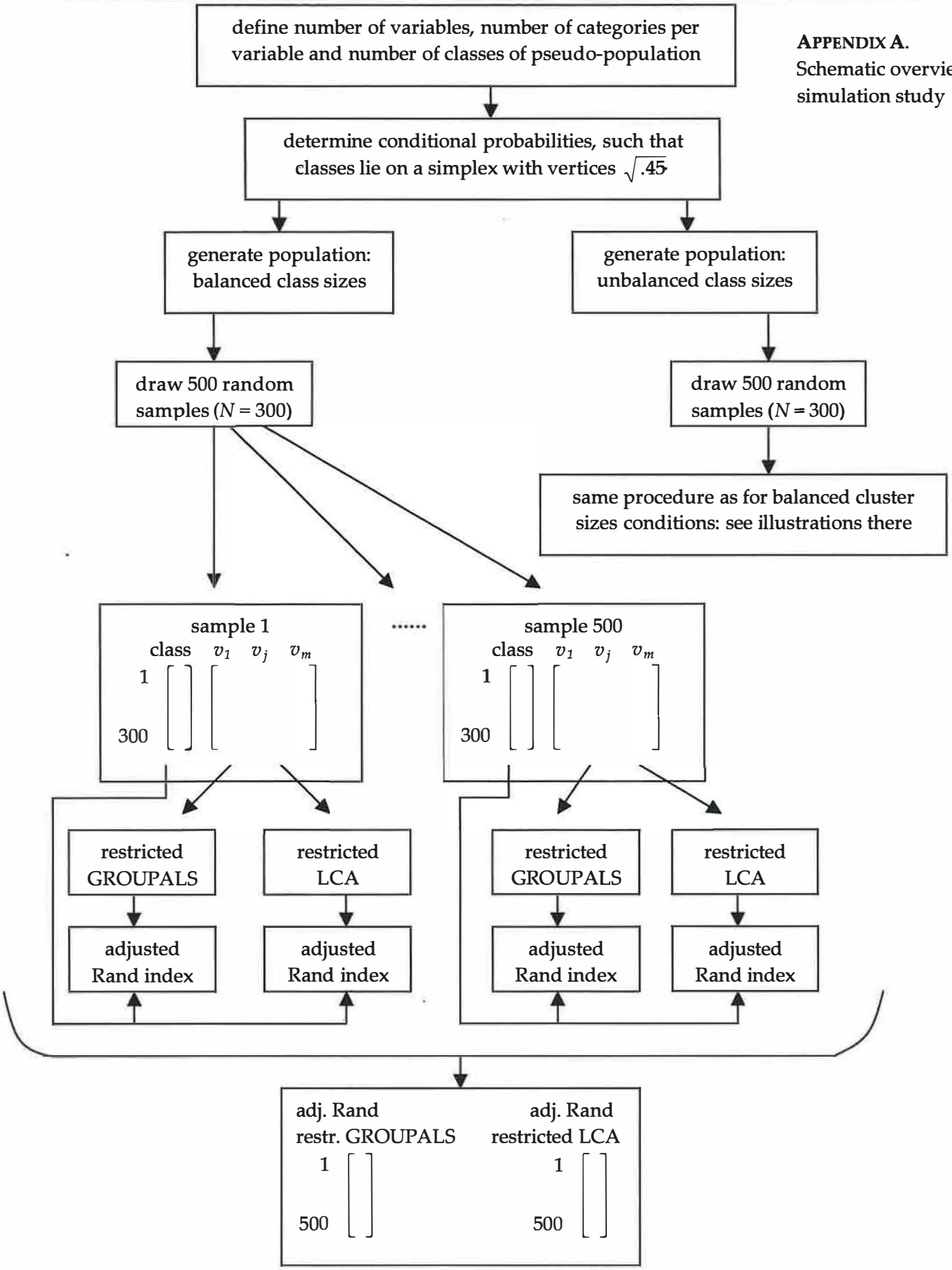
Van Os, B.J. (2000). *Dynamic Programming for Partitioning in Multivariate Data Analysis.* Leiden: Universal Press.

Van Putten, C.M., Van Den Brom-Snijders, P., & 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: Tilburg University.

Vermunt, J.K., & Magidson, J. (2002). Latent Class Cluster Analysis. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp.89 – 106). Cambridge: 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.



APPENDIX A. Schematic overview simulation study

APPENDIX B

Conditional probabilities used in data generation

3 classes, 5 categories

	cat1	cat2	cat3	cat4	cat5
α	.050	.100	.075	.297	.478
β	.602	.050	.100	.075	.173
γ	.200	.575	.050	.100	.075

3 classes, 7 categories

	cat1	cat2	cat3	cat4	cat5	cat6	cat7
α	.040	.050	.070	.100	.120	.055	.545
β	.519	.040	.050	.070	.100	.120	.101
γ	.077	.543	.040	.050	.070	.100	.120

4 classes, 5 categories

	cat1	cat 2	cat3	cat4	cat5
α	.050	.100	.075	.297	.478
β	.602	.050	.100	.075	.173
γ	.200	.575	.050	.100	.075
δ	.133	.138	.553	.125	.050

4 classes, 7 categories

	cat1	cat2	cat3	cat4	cat5	cat6	cat7
α	.050	.050	.070	.100	.020	.556	.154
β	.187	.050	.050	.070	.100	.020	.523
γ	.227	.483	.050	.050	.070	.100	.020
δ	.026	.183	.101	.553	.050	.070	.100

5 classes, 5 categories

	cat1	cat2	cat3	cat4	cat5
α	.050	.100	.095	.222	0.533
β	.592	.050	.100	.095	0.163
γ	.182	.573	.050	.100	0.095
δ	.117	.135	.549	.149	0.050
ε	.195	.138	.062	.602	0.001

5 classes, 7 categories

	cat1	cat2	cat3	cat4	cat5	cat6	cat7
α	.050	.050	.080	.100	.120	.034	.566
β	.512	.050	.050	.080	.100	.120	.088
γ	.058	.542	.050	.050	.080	.100	.120
δ	.073	.097	.550	.051	.050	.080	.100
ε	.051	.097	.130	.022	.570	.050	.080

APPENDIX C

Incidence of locally optimal solutions by restricted LCA

Table C. Proportion of samples for which adjusted Rand index has changed from first restricted LCA solution to final solution: impression of incidence of bad local optima

number of classes	relative class size	number of categories	5 variables	10 variables
3	**balanced**	**5**	.000	**.116**
		7	.000	**.000**
	unbalanced	**5**	.004	**.006**
		7	.004	**.006**
4	**balanced**	**5**	.004	**.004**
		7	.002	**.028**
	unbalanced	**5**	.092	**.104**
		7	.058	**.078**
5	**balanced**	**5**	.008	**.018**
		7	.024	**.038**
	unbalanced	**5**	.214	**.496**
		7	**.186**	**.294**