SOCIOLOGICAL METHODOLOGY

AMERICAN SOCIOLOGICAL ASSOCIATION

6

THE SIMULTANEOUS DECISION(S) ABOUT THE NUMBER OF LOWER- AND HIGHER-LEVEL CLASSES IN MULTILEVEL LATENT CLASS ANALYSIS

*Olga Lukočienė**

*Roberta Varriale**

*Jeroen K. Vermunt**

Recently, several types of extensions of the latent class (LC) model have been developed for the analysis of data sets having a multilevel structure. The most popular variant is the multilevel LC model with finite mixture distributions at multiple levels of a hierarchical structure; that is, with LCs for both lower-level units (e.g. individuals, citizens, or patients) and higher-level units (e.g. groups, regions, or hospitals). A problem in the application of this model is that determining the number of LCs is much more complicated than in standard (single-level) LC analysis because it involves multiple, nonindependent decisions. We propose a three-step model-fitting procedure for deciding about the number of higher- and lower-level classes. We also investigate the performance of information criteria (BIC, AIC, CAIC, and AIC3) in the context of multilevel LC analysis, with different types of response variables. A specific difficulty associated with using BIC and CAIC in any type of multilevel analysis is that these measures contain the sam-

Olga Lukočiené’s contribution to this research was funded by the Netherlands Organisation for Scientific Research, MaGW/NWO project number 400-03-295. Direct Correspondence to Jeroen K. Vermunt, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: J.K.Vermunt@uvt.nl.

*Tilburg University

ple size in their formulae, and we investigate whether this should be the number of groups, the number of individuals, or either the number of groups or individuals depending on whether one has to decide about model features concerning the higher or lower level. The three main conclusions of our simulations studies are that (1) the proposed three-step model-fitting strategy works rather well, (2) the number of higher-level units (K) is the preferred sample size for BIC and CAIC, both for decisions about higher- and lower-level classes, and (3) with categorical indicators, AIC3 and BIC based on the higher-level sample size are the preferred measures for deciding about the number of LCs at both the higher and lower level. With continuous indicators, BIC(K) performs better than AIC3. AIC performs best in very specific situations—namely, with poorly separated classes and categorical indicators.

1. INTRODUCTION

During recent decades, latent class (LC) analysis has become part of the standard statistical toolbox of researchers in applied areas such as medicine, biology, social sciences, psychology, education, criminology, and marketing. As is typical for most statistical techniques, one of the assumptions in LC modeling is that the available sample consists of a set of independent units, an assumption that is inadequate when units are nested within clusters sharing common environments, experiences, and interactions. In such situations, one should use multilevel techniques that take the dependencies between lower-level units resulting from the hierarchical data structure into account (Hox 2002; Snijders and Bosker 1999).

Recently, various types of multilevel extensions of LC and other types of finite mixture models have been developed (Asparouhov and Muthén 2008; Di and Bandeen-Roche 2008; Palardy and Vermunt forthcoming; Vermunt 2003, 2004, 2007, 2008). The common element of these extensions is that some of the LC model parameters are allowed to vary randomly across higher-level (group) units. Although group differences can also be modeled using multigroup LC analysis (Clogg and Goodman 1984), such an approach is feasible only when the number of groups is not too large, say between two and ten, because otherwise the number of parameters (one set for each group) becomes very large.

With larger numbers of (possibly small) groups, it is more appropriate to model group differences using random effects.

For example, in an LC analysis of a set of questions related to work satisfaction answered by employees of say 100 organizations, one has to take into account that work satisfaction may differ across organizations. This can be achieved by assuming that the class membership probabilities differ randomly across organizations, rather than estimating a different set of class membership probabilities for each organization. The multilevel LC model proposed by Vermunt (2003) involves expanding the standard (single-level) LC model with either a continuous or a discrete latent variable at the higher level, yielding either a parametric or a nonparametric specification for the random effects distribution (Aitkin 1999; Skrondal and Rabe-Hesketh 2004).

This paper deals with the nonparametric (or semiparametric) variant of the multilevel LC model in which differences across groups are modeled using a discrete latent variable at the group level. Applications of this variant of the multilevel LC model typically aim at simultaneously clustering individuals and groups; that is, lower-level units are assumed to belong to lower-level LCs differing in the distribution of the observed responses and higher-level units are assumed to belong to higher-level LCs differing in the distribution of the lower-level LCs. A good example is a recent study by Cavrinia, Galimberti, and Soffritti (2009) on patients' satisfaction with hospital services: the lower-level LCs represent clusters of patients with similar satisfaction levels concerning the studied aspects of hospital services, and LCs at the higher level represent clusters of hospitals with similar distributions of patients across the patient-level satisfaction clusters. Other applications of this variant of the multilevel LC model include studies by Bassi (2009), Bijmolt, Paas, and Vermunt (2004), Bouwmeester, Vermunt, and Sijtsma (2007), Henry and Muthén (forthcoming), Kragelj and Schlutter (2007), Pirani, Schifini, and Vermunt (2009), and Rindskopf (2006).

Even though the theory of multilevel LC analysis is well developed and interesting applications have already been published in a broad range of applied areas, one important issue has received little attention—namely, the problem related to the simultaneous decision regarding the number of lower- and higher-level LCs. For standard LC and standard mixture models, there is a large body of literature on the performance of statistics for determining the number of mixture components. It is well-known that asymptotic likelihood ratio tests cannot be used because certain regularity conditions do not hold, but that approximate *p*-values can be obtained using bootstrap procedures (McLachlan 1987; McLachlan and Peel 2000). However, because bootstrapping is computationally very intensive, applied researchers typically prefer using measures weighting model fit (the log-likelihood value) and model complexity (the number of parameters). The most popular of these measures is the Bayesian information criterion (BIC; Schwarz 1978; Hagenaars and McCutcheon 2002; Nylund et al. 2007). Other authors, however, suggest using the Akaike information criterion (AIC; Akaike 1974), at least in particular situations (Lin and Dayton 1997). Other alternatives are adjusted versions of AIC, such as consistent AIC (CAIC; Bozdogan 1987) and AIC3 (Bozdogan 1993).

Although deciding about the number of mixture components is already a rather complex task in standard LC and standard mixture modeling, it is even more complex in multilevel mixture modeling. It not only involves two decisions instead of one, about the number of both lower- and higher-level LCs, these decisions may also be mutually dependent. Except for the simulation study by Lukočienė and Vermunt (2010), this issue has not received any attention in the literature on multilevel LC analysis. However, these authors focused on the rather simplified situation in which the number of lower-level classes is known; that is, on the situation in which only one decision (about the higher-level classes) has to be made. Their simulation study showed that overall AIC3 performs best. Another important result is that the sample size in the BIC and CAIC formulas should be the number of higher-level units.

It is important to note that deciding about the number of mixture components is not always an issue in (multilevel) LC or mixture modeling. It is, of course, an issue when the model is used as a cluster technique with the aim of finding a good fitting and easy to interpret solution. However, mixture models can also be used as random effect models with a nonparametric specification of the random effects distribution (Aitkin 1999). In such applications, one should increase the number of LCs until the log-likelihood function reaches its maximum—that is, until the saturation point is reached where increasing the number of classes does no longer yield an increase of the log-likelihood.

This paper extends the work of Lukočienė and Vermunt (2010) in various ways. The most important extension is that it does not assume the number of lower-level LCs is known, but instead it deals with the situation encountered in practice in which both the number of higher- and lower-level LCs is unknown. In other words, we compare the performance of the most popular measures (BIC, AIC, CAIC, and AIC3) when simultaneously deciding about the number of mixture components at the lower and higher levels. A second extension is that we propose a stepwise model-fitting strategy that allows us to make the two decisions in a more efficient way. Moveover, the reported simulation studies look at a much larger range of conditions than those shown by Lukočienė and Vermunt (2010). Another extension is that we focus not only on LC models for categorical responses but also on models for continuous responses. Furthermore, contrary to Lukočienė and Vermunt (2010), the investigated approaches are illustrated using empirical applications. The first application concerns a multilevel LC analysis with a set of categorical indicators measuring the job satisfaction of graduates from different degree programs of the University of Florence, where the aim is to cluster both graduates and programs into homogeneous LCs. The second application deals with the analysis of a set of continuous intelligence measures taken from children nested within families and aims at clustering both children and their families.

Section 2 describes the multilevel LC model. The new three-step model-fitting procedure and the model-selection criteria that will be evaluated are described in Section 3. Sections 4 and 5 present the designs and the results of our two simulation studies dealing with categorical and continuous responses, respectively. Two applications are presented in Section 6, and Section 7 contains the main conclusions of our study.

2. THE MULTILEVEL LATENT CLASS MODEL

We denote the observed responses in a data set used to build a multilevel LC model by y_{kji}, where the indices i, j, and k refer to a response variable, an individual or lower-level unit, and a group or higher-level unit, respectively. The number of response variables equals $I (i = 1, \ldots, I)$, the number of individuals within group k equals $n_k (j = 1, \ldots, n_k)$, and the number of groups equals $K (k = 1, \ldots, K)$. Moreover, the total number of lower-level units equals $N = \sum_{k=1}^{K} n_k$. The vectors $\mathbf{y}_{kj} = (y_{kj1}, \ldots, y_{kji}, \ldots, y_{kjI})$ and $\mathbf{y}_k = (\mathbf{y}_{k1}, \ldots, \mathbf{y}_{kj}, \ldots, \mathbf{y}_{kn_k})$ contain the I responses of individual j from group k and the full set of responses of group k, respectively. Note that such a data set can be perceived as either an I-variate two-level data set or a univariate three-level data set.

A multilevel LC model assumes that individuals belong to one of L classes and that groups belong to one of H classes. The variables representing the lower- and higher-level class memberships are denoted by x_{kj} and w_k, respectively, and a particular class by $l(l = 1, \ldots, L)$ and $h(h = 1, \ldots, H)$, respectively.

The multilevel LC model proposed by Vermunt (2003, 2008) can be formulated using two basic equations. The first equation defines the (mixture) model for $f(\boldsymbol{y}_k)$, the marginal density of the full response vector of group k; that is,

$$f(\boldsymbol{y}_k) = \sum_{h=1}^{H} P(w_k = h) \prod_{j=1}^{n_k} f(\boldsymbol{y}_{kj}|w_k = h). \tag{1}$$

Here, $P(w_k = h)$ is the probability that group k belongs to LC h and that $f(\boldsymbol{y}_{kj}|w_k = h)$ is the conditional density for the response vector of individual j in group k conditional on the membership of group k to LC h. The second equation defines the (mixture) model for $f(\boldsymbol{y}_{kj}|w_k = h)$; that is,

$$f(\boldsymbol{y}_{kj}|w_k = h) = \sum_{l=1}^{L} P(x_{kj} = l|w_k = h) \prod_{i=1}^{I} f(y_{kji}|x_{kj} = l, w_k = h), \tag{2}$$

where $P(x_{kj} = l|w_k = h)$ is the probability that individual j of group k belongs to LC l given that the group belongs to LC h, and $f(y_{kji}|x_{kj} = l, w_k = h)$ is the conditional density for response variable i of individual j in group k given the membership to individual-level class l and group-level class h.

These two equations clearly show which conditional independence assumptions are made in a multilevel LC analysis. First, the observations of the n_k individuals in group k are assumed to be independent of one another given the group-level class membership. Note that this assumption is typical for any type of multilevel analysis: observations are assumed to be independent conditional on the random effects (Skrondal and Rabe-Hesketh 2004). Second, the I responses of individual j are assumed to be independent of each other given the group and individual LC memberships, which is the basic assumption of most LC models and is usually referred to as the local independence assumption (Bartholomew and Knott 1999; Hagenaars and McCutcheon 2002).

The last element in a multilevel LC model is the specification of the conditional densities $f(y_{kji}|x_{kj} = l, w_k = h)$, which will typically be assumed to belong to the exponential family. This can, for example, be a normal or gamma distribution for continuous responses, a Poisson, binomial, or negative binomial distribution for counts, and a multinomial distribution for categorical responses. In the current paper, we restrict ourselves to models for either categorical or continuous responses. In models for categorical responses, $y_{kji} = 1, \ldots, M_i$, where M_i is the number of categories of the ith response variable, and the multinomial form of density $f(y_{kji}|x_{kj} = l, w_k = h)$ can be expressed as

$$f(y_{kji}|x_{kj} = l, w_k = h) = \prod_{m=1}^{M_i} (\pi_{hlim})^{d_{kjim}}, \tag{3}$$

where d_{kjim} represents an indicator variable taking on the value 1 if $y_{kji} = m$ and 0 otherwise, and where π_{hlim} represents a multinomial probability subject to the constraints $\pi_{hlim} \geq 0$ and $\sum_{m=1}^{M_i} \pi_{hlim} = 1$. Continuous responses are typically assumed to come from normal distributions with class-specific means and variances; that is, $f(y_{kji}|x_{kj} = l, w_k = h) \sim N(\mu_{hl}, \sigma^2_{hl})$.

As standard single-level LC models, the multilevel LC model can easily be extended to include explanatory variables affecting the responses and the lower- and higher-level class memberships. This involves conditioning the response variables' densities and the class membership probabilities on covariates (for some examples, see Vermunt 2003, 2004, 2008). Here, we restrict ourselves to models without covariates.

Equations (1) and (2) describe the multilevel LC model (without explanatory variables) in its most general form; that is, as a model in which both the lower-level mixture proportions—$P(x_{kj} = l|w_k = h)$—and the parameters defining the response densities—$f(y_{kji}|x_{kj} = l, w_k = h)$—are allowed to differ across higher-level classes. The only application of this general model we know about is the one described by Henry and Muthén (forthcoming). Most applications of multilevel LC analysis, however, use one of two more restricted special cases. More specifically, they impose one of the following two constraints:

(1) $f(y_{kji}|x_{kj} = l, w_k = h) = f(y_{kji}|x_{kj} = l)$; or
(2) $P(x_{kj} = l|w_k = h) = P(x_{kj} = l)$.

In the first restricted special case, $P(x_{kj} = l|w_k = h)$ is estimated freely, but the parameters defining the conditional distributions are assumed to be independent of the higher-level class membership (Vermunt 2003, 2008). This structure is the one used in almost all the applications listed in the introduction—that is, in applications aiming at the simultaneous clustering of higher- and lower-level units. In fact, the clustering of higher-level units is performed by "pushing up" the information contained in the multiple lower-level responses via the lower-level class memberships.

In the second special case, the parameters defining $f(y_{kji}|x_{kj} = l, w_k = h)$ are estimated freely, but the lower-level class membership is assumed to be independent of the higher-level class membership (Vermunt 2004). This specification is in fact similar to the variance decomposition used in three-level regression models: the variation in the responses is split into between-group and within-group parts (Skrondal and Rabe-Hesketh 2004). In our simulation studies, we focus on the first specification, which is the one that has been used in almost all applications of multilevel LC analysis that have been published so far (see also the introduction).

3. DETERMINING THE NUMBER OF LOWER- AND HIGHER-LEVEL CLASSES

3.1. *A Three-Step Model-Fitting Procedure*

Determining the number of classes in multilevel LC analysis involves a simultaneous decision regarding the number of LCs at multiple levels of the hierarchical structure. The main complication is that these decisions are not mutually independent. In a model with a structure corresponding to the first special case discussed above, the higher-level classes differ only with respect to their lower-level class distributions. It is therefore not surprising that the selected number of higher-level classes depends very much on the selected number of lower-level classes. Although the reversed dependency will typically be less strong, it may also exist, especially when (some of) the lower-level classes are only weakly defined (when classes are not very well separated). In such situations, the multilevel data structure with observations that are dependent within groups may yield important additional information on the lower-level class memberships; that is, the responses of the other group members may be informative about a person's own class membership. This mechanism requires that the dependencies are picked up by the higher-level classes.

The model–fitting strategy used in the first paper on multilevel LC analysis (Vermunt 2003)—and which is also the strategy used in most applications of this model—is in fact a two-step procedure. We first determine the number of lower-level classes ignoring the multilevel structure and subsequently determine the number of higher-level classes fixing the number of lower-level classes at the value from the first step. It should be noted that the simulation study by Lukočiené and Vermunt (2010) on the selection of the number of higher-level classes builds on this model selection strategy in that it investigates the performance of various model selection criteria in the second step. The main disadvantage of this two-step strategy is that it accounts only partially for the dependency between the two decisions to be made. More specifically, the dependency of the decision about the number of lower-level classes on the selected number of higher-level classes is fully ignored.

Bijmolt and colleagues (2004) used an alternative model-fitting strategy that involves estimating the multilevel LC model for all relevant combinations of L and H. In their application, this implied estimating models with L ranging from 1 to 15 and H ranging from 1 to 8. Vermunt (2008) used the same procedure in a set of applications illustrating the use of multilevel LC analysis in medical research. The two main disadvantages of this procedure are that it may require estimating a large number of models (more than 100 in the Bijmolt et al. application) and that it does not allow the use of different measures when deciding about the value of L and H (see also below).

We propose an alternative three-step model-fitting procedure that (1) is less computationally intensive than the procedure by Bijmolt and colleagues (2004), (2) accounts for the fact that the value of L may depend on the selected value of H, and (3) allows the use of different measures when deciding about L and H. This procedure consists of three steps:

1. Determine the number of lower-level classes ignoring the multilevel structure (that is, assuming that $H = 1$).
2. Fix the number of lower-level classes to the value of step 1 and determine the number of higher-level classes.
3. Fix the number of higher-level classes to the value of step 2 and redetermine the number of lower-level classes.

Note that the first two steps are the same as the ones used by Vermunt (2003) but with the important modification that different fit indices may be used in steps 1 and 2 (more details are provided below). The aim of the extra step 3 is to evaluate whether the number of lower-level classes changes after taking into account the dependencies between lower-level units due to the multilevel data structure. Of course, a fourth step could be added in which the number of higher-level classes is reevaluated fixing L to the value of step 3, as well as a fifth step in which the number of lower-level classes is reevaluated fixing H to the value of step 4, etc. In the current study, however, we restrict ourselves to the above three-step approach, which we believe already provides an important improvement compared to the approaches offered by Vermunt (2003) and Bijmolt and colleagues (2004).

3.2. *Model Selection Measures*

When working within a maximum likelihood estimation framework as we do here, comparison of nested models is typically performed by means of likelihood-ratio tests, which under certain regularity conditions follow a chi-squared distribution. However, such likelihood-ratio tests cannot be used to compare models with different numbers of classes because the null model with the smaller number of classes is obtained by fixing one or more parameters of the alternative model at their boundary values. A solution proposed by various authors is to use parametric bootstrap procedures to approximate the p-value associated with these likelihood-ratio tests (for example, see McLachlan 1987; Nylund et al. 2007). However, these bootstrap-based testing procedures are seldom used by applied researchers because they are computationally very intensive and, moreover, their correct implementation is not at all straightforward.

Most researchers applying LC analysis will make use of information criteria that are measures weighting model fit (the log-likelihood value) and model complexity (the number of parameters). As the log-likelihood will typically increase (until its saturation point) with increasing model complexity (with increasing number of classes), it is penalized by the addition of a term measuring the complexity of the model. These information criteria can be expressed most generally as

$$IC = -2 \log L(\boldsymbol{\theta}) + Cr, \tag{4}$$

where $L(\boldsymbol{\theta})$ is the maximized log-likelihood value for a model with parameters $\boldsymbol{\theta}$, r is the number of independent parameters in this model, and C is the weight given to the penalty term based on r. The lower the value of an information criterion, the better the model. The various information criteria proposed in the literature differ in the value of C.

Most texts on LC analysis suggest using the Bayesian information criterion (BIC; Schwarz 1978) for deciding about the number of classes (for example, see Hagenaars and McCutcheon 2002; Magidson and Vermunt 2004). BIC is defined as

$$BIC = -2 \log L(\boldsymbol{\theta}) + \log(n)r, \tag{5}$$

where n is the number of observations (sample size). Simulation studies have shown that BIC usually performs very well but also that it may sometimes underestimate the number of classes, especially when classes are not well separated (for example, see Dias 2004; Nylund et al. 2007).

Others suggest using the Akaike information criterion (AIC; Akaike 1974), which is expressed as

$$AIC = -2 \log L(\boldsymbol{\theta}) + 2r. \tag{6}$$

Simulation studies have shown that AIC tends to overestimate the number of classes (McLachlan and Peel 2000; Dias 2004), although others report that AIC works well in specific situations (Lin and Dayton 1997).

Bozdogan proposed two adjusted versions of AIC—AIC3 (Bozdogan 1993) and consistent AIC (CAIC; Bozdogan 1987)—which are used more and more in LC analysis. AIC3 and CAIC can be expressed, respectively, by

$$AIC3 = -2\log L(\boldsymbol{\theta}) + 3r \tag{7}$$

and

$$CAIC = -2\log L(\boldsymbol{\theta}) + (1 + \log(n))r. \tag{8}$$

Simulation studies by Andrews and Currim (2003) and Dias (2004) showed that AIC3 is the best-performing criterion in LC analysis with categorical response variables. Note that the AIC3 weight of 3 typically falls between the BIC weight of log n and the AIC weight of 2. It can thus be seen as a compromise between these two measures that, compared to BIC, is better able to detect poorly separated classes and that, contrary to AIC, is less likely to come up with spurious classes. The reported behavior of CAIC is similar to the behavior of BIC, which is not surprising given that their penalties are rather similar.

There is a large number of simulation studies on the performance of AIC, AIC3, and BIC, as well as related likelihood-based measures, in the context of mixture models for continuous response variables (see, among others, Bezdek, Attikiouzel, and Windham 1997; Biernacki 1997; Biernacki, Celeux, and Govaert 2000; Bozdogan 1994; Cutler and Windham 1994; McLachlan and Peel 2000, ch. 6; Fraley and Raftery 1998). Fonseca and Cardoso (2007) provided an overview of these studies and those focusing on categorical responses, and they concluded that AIC3 works best with categorical responses and BIC with continuous responses.

Lukočienė and Vermunt (2010) pointed out a specific issue when using BIC and CAIC in the context of multilevel analysis: It is not clear whether the sample size should be the number of groups (K), the total number of individuals (N), or either the number of groups or number of individuals depending on whether one wishes to test model features related to the higher or lower level. Work by Pauler (1998) on the use of BIC in the context of univariate linear mixed models suggests that one should use K for decisions about higher-level model features and N for lower-level model features.

The aim of the current study is to determine the performance of the various information criteria described above for deciding about the number of classes in multilevel LC models. That is, the question is whether results found for single-level LC models for categorical responses and mixture models for continuous responses also apply to multilevel generalizations of these models. The work by Lukočiené and Vermunt (2010) is the only study that has been published on this topic, but it restricted itself to the simplified situation in which the number of lower-level classes can be assumed to be known. The results of this study can be assumed to be valid in step 2 of the three-step model-fitting procedure described above, but only if L was correctly estimated in step 1. There are two main results in the Lukočiené and Vermunt (2010) study: (1) that K should be used as the sample size in the BIC and CAIC formulas when deciding about the number of higher-level classes, and (2) that overall, as in standard single-level LC models, AIC3 is the preferred measure.

The current study aims at providing information on the performance of the various information criteria in the more realistic situation in which the number of lower-level LCs is unknown. We will again address the issue related to sample size definition in BIC and CAIC, but now for the selection of not only the number of higher-level classes but also the number of lower-level classes. Moreover, we will investigate the possibility of using different sample size definitions in steps 1 and 3 on the one hand and step 2 on the other hand. Another departure from the work of Lukočiené and Vermunt (2010) is that we investigate LC models not only for categorical indicators but also for continuous indicators.

4. DESIGN OF THE SIMULATION STUDIES

There are two main questions addressed in the simulation studies:

1. How well does the proposed three-step model-fitting procedure perform under the studied conditions?
2. How well do the various information criteria perform under the studied conditions?

By performance we mean whether the model with the correct number of LCs is selected by our procedure. The starting point for the design of the simulation studies—for defining the conditions that will be varied—is what is known from previous simulation studies on determining the number of classes in LC models. As summarized by Dias (2004), there are two main factors determining the difficulty of detecting the correct number of classes:

1. the separation between the classes (the smaller the separation between the classes the less likely that one finds the right number of classes), and
2. the sample size (the smaller the sample size the less likely that one finds the right number of classes).

These are the two key factors that will be manipulated, and because we are dealing with a multilevel LC model instead of a standard LC model, these will be manipulated for both the higher and the lower level.

It should be noted that while "separation between the classes" has been reported to be the most important factor (for example, see Andrews and Currim 2003; Dias 2004; Sarstedt 2008), it is also a somewhat "obscure" factor because it can be manipulated and quantified in various ways. As is often done in LC and mixture modeling, we will quantify the separation between classes using an entropy-based R-squared measure indicating how well the class memberships can be predicted from the observed responses (Wedel and Kamakura 1998). For the lower level, this measure can be defined as

$$R^2_{entropy,low} = 1 - \frac{\sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{l=1}^{L} -P(x_{kj} = l|\mathbf{y}_k)\log P(x_{kj} = l|\mathbf{y}_k)}{\sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{l=1}^{L} -P(x_{kj} = l)\log P(x_{kj} = l)}, \tag{9}$$

and for the higher level as

$$R^2_{entropy,high} = 1 - \frac{\sum_{k=1}^{K}\sum_{h=1}^{H} -P(w_k = h|\mathbf{y}_k)\log P(w_k = h|\mathbf{y}_k)}{\sum_{k=1}^{K}\sum_{h=1}^{H} -P(w_k = h)\log P(w_k = h)}. \tag{10}$$

Note that these measures quantify the relative improvement of the prediction of the class membership when using the responses (conditional entropy in the numerator) compared to the prediction without using the responses (unconditional entropy in the denominator). A value equal to 0 corresponds to a prediction that is no better than chance (and thus no separation at all) and a value of 1 to a perfect prediction (and thus a perfect separation).

The $R^2_{entropy,low}$ depends on the number of lower-level classes, the number of response variables, and the parameters defining the class-specific response densities. It will be larger with a smaller number of classes, a larger number of response variables, and larger between-class and smaller within-class variation in responses. For categorical responses, the latter corresponds to a larger number of categories and larger differences in the class-specific response probabilities, and, for continuous responses, to larger differences in the class-specific means and smaller within-class variances. The $R^2_{entropy,high}$ depends on the number of classes at the higher level, the number of individuals per group, the number of classes at the lower level, and the conditional probabilities $P(x_{kj} = l|w_k = h)$. The separation at the higher level is larger with a smaller number of higher-level classes, a larger number of individuals per group, a larger number of lower-level classes, and larger differences in the conditional probabilities $P(x_{kj} = l|w_k = h)$ across higher-level classes. Note that the number of lower-level classes affects the entropy at both levels, but in an opposite direction.

The settings for the design factors we used in our simulation studies have been chosen in order to cover a broad range of possible separation values at both levels. We describe below the design of our two simulation studies in more detail. The first study focuses on LC models for categorical indicators and the second on LC models for continuous indicators.

4.1. *Study I: Categorical Indicators*

The first simulation study concerns multilevel LC analysis with categorical indicators. The following design factors were varied:

1. the number of lower-level classes L,
2. the number of higher-level classes H,
3. the lower-level class probabilities $P(x_{kj} = l|w_k = h)$,
4. the number of response variables I,

5. the class-specific response probabilities π_{lim}
6. the lower-level sample size n_k, and
7. the higher-level sample size K.

Two factors were constant across simulation replications: (1) the higher-level class sizes, which were always set to $P(w_k = h) = 1/H$, and (2) the number of categories of the response variables, which were fixed to $M_i = 2$. The latter implies that our simulation study concerns multilevel LC models for dichotomous responses. It should be noted that by varying M_i we would primarily introduce another factor affecting the lower-level separation, which is already varied by means of the choice of L, I, and π_{lim}.

The factors related to the number of classes are the easiest to manipulate. More specifically, the number of LCs are either two or three at both levels ($L = 2, 3$ and $H = 2, 3$). Table 1 presents the structure used for varying the conditional class probabilities $P(x_{kj} = l|w_k = h)$. These are such that only one parameter (denoted by p_x) needs to be specified. We used either $p_x = 0.7$ or $p_x = 0.8$, where the larger value corresponds to more diverse conditional lower-level class probabilities and thus to better higher-level separation.

The number of items I was either 6 or 10. For the class-specific response probabilities π_{lim}—which are assumed to be unrelated to the higher-level class membership (the first restricted special case discussed in Section 2)—we used three settings that were defined with a single parameter denoted by p. More specifically, p could take on the values 0.7, 0.8, and 0.9. This parameter represents the probability of the first response for all items in the first class ($\pi_{1i1} = p$) and the probability of the second response in the last class ($\pi_{Li2} = p$). In the $L = 3$ condition, for class 2, it represents the probability of the first response for the first $I/2$ items and the probability of the second response for the remaining items. The interpretation of the classes is such that the first and last class are opposites, and in the $L = 3$ condition, the second class is similar to class 1 for half of the items and to class 3 for the other half. Note that $p = 0.7$ yields the smallest and $p = 0.9$ the largest separation between the lower-level classes.

Another factor that we manipulated is the number of lower-level units per group (the lower-level sample size n_k). Note that the larger the number of units per group, the more information we have about the group-level class membership. More specifically, we used the values

TABLE 1
Assumed Values for the Lower-Level LC Probabilities Conditional on the Higher-Level Class—($P(x_{kj} = l|w_k = h)$)—for $L = 2, 3$ and $H = 2, 3$

(a)

	h	
l	1	2
1	p_x	$1 - p_x$
2	$1 - p_x$	p_x

(b)

		h	
l	1	2	3
1	p_x	0.5	$1 - p_x$
2	$1 - p_x$	0.5	p_x

(c)

	h	
l	1	2
1	p_x	$(1 - p_x)/2$
2	$(1 - p_x)/2$	$(1 - p_x)/2$
3	$(1 - p_x)/2$	p_x

(d)

		h	
l	1	2	3
1	$(1 - p_x)/2$	0.33	p_x
2	$(1 - p_x)/2$	0.33	$(1 - p_x)/2$
3	p_x	0.33	$(1 - p_x)/2$

$n_k = 5, 10, 20$, and 50 to create conditions ranging from very low to very high separation.

The last factor that was varied is the higher-level sample size, for which we used $K = 30, 100$, and 1000. These sample sizes were chosen to cover the full range of small, moderate, and large sample sizes encountered in multilevel applications in social science research.

Table 2 presents the lowest and average $R^2_{entropy,low}$ and $R^2_{entropy,high}$ value for each of the manipulated conditions (the highest value is always close to 1). It can be seen that for the lower level the separation indeed depends on L, p, and I, and for the higher level

TABLE 2
Entropy-Based R-Squared Values at Lower and Higher Levels for all Values of n_k, K, H, L, p_x, p, and I, and for the $R^2_{entropy}$ Quintiles

		Higher Level		Lower Level	
		Lowest	Average	Lowest	Average
n_k	5	0.13	0.50	0.38	0.78
	10	0.22	0.68	0.40	0.79
	20	0.37	0.83	0.41	0.79
	50	0.62	0.95	0.42	0.80
K	30	0.13	0.74	0.38	0.79
	100	0.13	0.74	0.38	0.79
	1000	0.13	0.74	0.38	0.79
H	2	0.28	0.86	0.42	0.80
	3	0.13	0.63	0.38	0.78
L	2	0.13	0.66	0.51	0.84
	3	0.25	0.82	0.38	0.75
p_x	0.7	0.13	0.67	0.38	0.78
	0.8	0.26	0.81	0.42	0.80
p	0.7	0.13	0.68	0.38	0.58
	0.8	0.18	0.76	0.66	0.83
	0.9	0.20	0.79	0.89	0.96
I	6	0.13	0.73	0.38	0.74
	10	0.16	0.76	0.53	0.84
$R^2_{entropy}$ quintile	1	0.13	0.34	0.38	0.51
	2	0.48	0.61	0.59	0.68
	3	0.73	0.80	0.75	0.83
	4	0.90	0.95	0.90	0.94
	5	0.95	1.00	0.97	0.98
Total		0.13	0.74	0.38	0.79

mainly on n_k, H, L, and p_x, but also slightly on p, and I (indirect on the lower-level separation). These numbers show that our settings are such that we cover a broad range of separation values both for the lower- and higher-level classes.

In total, the simulation study design contained $2 \times 2 \times 2 \times 2 \times 3 \times 4 \times 3 = 576$ cells representing all possible combinations of the seven varied design factors. For each of these cells, we generated five data sets, and ran models with different numbers of lower- and higher-level classes. The syntax version of Latent GOLD (Vermunt and Magidson 2008) was used for the realization of this simulation study, as well as for the study described next and the applications.

4.2. *Study II: Continuous Indicators*

In the second simulation study, we dealt with multilevel LC models with continuous indicators. The aim of this study was to check whether results depend on the type of response variables used in the multilevel LC model. As mentioned earlier, there is a large body of literature on deciding about the number of mixture components in models for continuous responses (for example, see Bezdek et al. 1997; Biernacki 1997; Biernacki et al. 2000; Bozdogan 1994; Cutler and Windham 1994; McLachlan and Peel 2000, ch. 6; Fraley and Raftery 1998). However, these studies focused on single-level mixture models, whereas here we are interested in the performance of information criteria in multilevel mixture modeling.

For the design factors H, L, p_x, n_k, and K, we used the same settings as in the first study. To reduce the size of this second study, we kept the number of items fixed to 6. This means that we should compare the results with the ones for the $I = 6$ condition in Study I.

The class-specific response densities are now defined by the item means and variances. The variances were all set to 1 ($\sigma_{li}^2 = 1$). Similar to π_{lim} in Study I, we used three settings for the means μ_{li}, defined by a single parameter d taking on the value 0.4, 0.7, or 1.0. The first class had item means equal to $-d(\mu_{1i} = -d)$, the last class to $d(\mu_{Li} = d)$, and in the $L = 3$ condition, the second class has $\mu_{2i} = -d$ for the first $I/2$ items and $\mu_{2i} = d$ for the rest. These three settings for d yielded very similar lower-level entropy values as the three settings for p in simulation Study I with $I = 6$. The average lower-level entropy is 0.74 in Study II, which is the same as the value for $I = 6$ in Study I. The average higher-level entropy is slightly lower in Study II (0.72 instead of 0.73).

5. RESULTS OF THE SIMULATION STUDIES

5.1. *Study I: Categorical Indicators*

The results obtained with simulation Study I, which deals with multilevel LC models for categorical responses, are presented below. We discuss the results for the lower-level classes, for the higher-level classes, and the overall results concerning the simultaneous decision about L and H. We also note the effect of using the third step of our stepwise modeling procedure.

5.1.1. *Lower-Level Classes*

Table 3 presents the results for the lower-level classes obtained after step 3. Per design factor and information criterion, it reports the percentage of simulation replications in which the number of classes is underestimated ($\hat{L} < L$), correctly estimated ($\hat{L} = L$), and overestimated ($\hat{L} > L$).

The lower-level results are very much in agreement with the results of simulation studies for standard (single-level) LC models (Andrews and Currim 2003; Dias 2004; Sarstedt 2008). Indeed, sample size, number of classes, number of items, and size of the response probabilities affect the difficulty of finding the correct model in the expected direction. Moreover, AIC3 is the best-performing criterion. BIC(N) and CAIC(N) are more likely than AIC3 to underestimate the number of classes with smaller sample sizes and lower separation levels. Moreover, AIC is more likely to overestimate the number of classes in all situations.

Comparison of the performance of the somewhat unconventional BIC(K) and CAIC(K) measures with the AIC3, BIC(N), and CAIC(N) shows that these perform almost as well as AIC3, and thus better than BIC(N) and CAIC(N). It should be noted that under all conditions the weight log K is closer to the AIC3 weight of 3 than log N, which is probably why the BIC(K) and CAIC(K) results are in line with the comparatively good qualities of AIC3.

Instead of looking at the separate effects of L, I, and p, we can also look at the overall effect of the lower-level entropy on the performance of the various measures. Table 3 provides the results for the five lower-level entropy quantiles. As can be seen, the higher the entropy values, the better the performance of the information criteria, especially

TABLE 3
Percentage of Simulation Replications in Which the Number of Lower-Level Classes is Underestimated, Correctly Estimated, and Overestimated

		n_k				K			H		L		p_x		p			I		$R^2_{entropy,low}$ quintile					
		5	10	20	50	30	100	1000	2	3	2	3	0.7	0.8	0.7	0.8	0.9	6	10	1	2	3	4	5	Total
BIC(N)	$\hat{L} < L$	22	16	12	7	26	15	2	16	12	1	27	13	15	31	10	1	19	9	45	21	4	1	0	14
	$\hat{L} = L$	78	84	88	93	74	85	98	84	88	99	73	87	85	69	90	99	81	91	55	79	96	99	100	86
BIC(K)	$\hat{L} < L$	18	12	7	3	17	11	1	12	8	0	19	9	11	24	5	0	14	6	35	13	1	0	0	10
	$\hat{L} = L$	83	88	93	97	83	89	99	88	92	100	81	91	89	76	95	100	86	94	65	87	99	100	100	90
CAIC(N)	$\hat{L} < L$	18	14	12	7	25	13	1	15	11	0	25	12	14	29	9	1	17	9	41	19	4	1	0	13
	$\hat{L} = L$	82	86	88	93	75	87	99	85	89	100	75	88	86	71	91	99	83	91	59	81	96	99	100	87
CAIC(K)	$\hat{L} < L$	20	14	9	4	21	13	2	13	10	1	23	11	13	28	7	0	16	7	41	16	2	1	0	12
	$\hat{L} = L$	80	86	91	96	79	87	98	87	90	99	77	89	87	72	93	100	84	93	59	84	98	99	100	88
AIC	$\hat{L} < L$	8	4	2	1	8	3	0	5	3	0	7	3	4	10	1	0	7	1	15	4	0	0	0	4
	$\hat{L} = L$	70	71	67	70	69	68	71	70	69	75	63	67	71	61	70	78	91	48	67	65	73	67	74	69
	$\hat{L} > L$	22	25	31	30	23	29	29	26	28	25	29	29	25	29	29	23	3	51	18	31	27	33	26	27
AIC3	$\hat{L} < L$	14	9	5	2	16	7	0	9	6	0	15	7	8	18	4	0	11	4	27	10	1	0	0	7
	$\hat{L} = L$	86	90	94	98	84	93	99	91	93	100	84	93	91	81	95	100	89	95	73	89	99	99	100	92
	$\hat{L} > L$	0	0	1	1	1	1	1	0	1	0	1	1	0	1	1	0	0	1	0	1	1	0	0	1

when comparing the first and second quintile as well as the second and the third. The only exception is the AIC, which does not seem to be affected by the entropy value in a systematic way.

5.1.2. *Higher-Level Classes*

Table 4 presents the results for the higher level classes obtained after step 2. Per design factor and information criterion, it reports the percentage of simulation replications in which the number of classes is underestimated ($\hat{H} < H$), estimated correctly ($\hat{H} = H$), and overestimated ($\hat{H} > H$).

As mentioned earlier, the key factors expected to affect the performance of the various information criteria are sample size and separation between classes. For the higher level, the sample size is defined by K and separation depends most strongly on H, p_x, and n_k, and somewhat on L, I, and p. The results of Table 4 show that each of the investigated criteria performs better under the easier conditions (larger sample and larger separation between classes). Another thing that can be observed is that the hit rates are lower than for the lower-level part of the model (see again Table 3). The explanation for this is that the separation values used for the higher level were slightly lower than the ones for the lower level (as shown in Table 2, the average entropy-based R-squared is .74 for the higher level and .79 for the lower level).

Comparing the various measures with one another shows that overall AIC3 and BIC(K) perform better than the other measures. The main difference between these two is that AIC3 performs slightly better than BIC(K) with lower separation (smaller n_k and lower quintiles of R-squared) and BIC(K) somewhat better than AIC3 for the high separation levels.

Both for BIC and CAIC, we find substantial differences between the versions based on sample size K and N. BIC(N) and CAIC(N) are more likely to underestimate the number of mixture components than BIC(K) and CAIC(K). Overall, CAIC performs slightly worse than BIC. Finally, we see again that AIC has the tendency to overestimate the number of LCs, but also that it performs better than the other methods under the lowest separation conditions.

As we saw for the lower-level results, the larger the separation between the LCs, the better the indices detect the correct number of LCs. This positive relationship does, however, not apply for BIC(K), AIC3, and AIC when going from the fourth to the fifth entropy quintile,

TABLE 4
Percentage of Simulation Replications in Which the Number of Higher-Level Classes is Underestimated, Correctly Estimated, and Overestimated

		n_k				K			H		L		p_x		p			I		$R^2_{entropy,high}$ quintile					
		5	10	20	50	30	100	1000	2	3	2	3	0.7	0.8	0.7	0.8	0.9	6	10	1	2	3	4	5	Total
BIC(N)	$\hat{H} < H$	53	37	22	6	43	31	15	5	54	38	21	36	23	38	27	24	32	28	85	47	15	1	0	30
	$\hat{H} = H$	47	63	78	94	57	69	85	95	46	62	79	64	77	62	73	76	68	72	15	53	85	99	100	70
BIC(K)	$\hat{H} < H$	49	33	16	3	36	26	14	3	47	33	17	31	19	33	23	20	27	23	82	39	6	0	0	25
	$\hat{H} = H$	51	67	84	96	64	74	86	97	53	66	83	69	81	68	77	80	73	77	18	61	94	100	99	75
	$\hat{H} > H$	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CAIC(N)	$\hat{H} < H$	49	36	21	8	43	29	13	4	53	37	20	35	22	37	26	23	31	26	81	44	16	2	0	28
	$\hat{H} = H$	51	64	79	92	57	71	87	96	47	63	80	65	78	63	74	77	69	74	19	56	84	98	100	72
CAIC(K)	$\hat{H} < H$	51	35	18	4	38	28	14	4	50	35	19	33	21	34	24	22	29	25	84	43	8	0	0	27
	$\hat{H} = H$	49	65	83	96	62	72	86	96	50	65	81	67	79	66	75	78	71	75	17	57	92	100	100	73
AIC	$\hat{H} < H$	40	25	10	2	30	19	9	1	37	27	12	24	15	24	17	17	21	17	67	27	3	0	0	19
	$\hat{H} = H$	56	67	75	82	63	69	78	85	55	66	75	66	74	66	72	73	72	68	31	67	85	87	80	70
	$\hat{H} > H$	4	8	14	16	7	11	14	13	8	8	14	10	11	10	11	11	7	15	1	7	12	13	20	11
AIC3	$\hat{H} < H$	46	30	14	3	35	24	11	3	44	31	15	29	18	29	21	19	25	21	77	34	6	0	0	23
	$\hat{H} = H$	54	69	85	93	64	75	87	95	55	68	83	70	80	69	77	79	74	76	23	65	93	97	96	75
	$\hat{H} > H$	1	1	2	4	1	2	2	2	1	1	3	1	2	2	2	2	1	2	0	0	1	3	4	2

although the hit rate is still very high (larger than 95 percent) in the fifth quintile.

5.1.3. *Combined Lower- and Higher-Level Classes*

The main goal of this simulation study was to determine which of the investigated model selection measures is preferable for deciding simultaneously about the number of lower- and higher-level classes in multilevel LC models. Tables 5 and 6 present the percentage of simulation replications in which the number of lower- and higher-level classes is correctly estimated ($\hat{L} = L$ and $\hat{H} = H$) for the design factors and the separation quintiles, respectively. Note that we also present the results for BIC and CAIC with sample size N for the lower-level analysis (steps 1 and 3) and K for the higher-level analysis (step 2), denoted by BIC(N, K) and CAIC(N, K).

Comparison of the results for the investigated fit measures shows that overall AIC3 and BIC(K) have higher hit rates than the other indices, except for the lowest higher-level entropy conditions, for which AIC performs better. Comparison of AIC3 with BIC(K) shows that the latter performs better with well-separated lower- and higher-level classes and the former with more poorly separated classes. BIC(K) performs slightly better than BIC(N, K) and better than BIC(N), and the same applies to the three versions of CAIC. However, BIC performs better than CAIC with the same sample size definitions.

As can be clearly seen from Table 6, for each measure applies that the better separated the lower- and higher-level classes are, the better the hit rates. An exception is the fifth lower-level quintile for which the hit rates are lower than for the fourth quintile. This is probably due to the fact that the fifth quintile mainly contains design cells with $L = 2$, and the higher-level entropy is somewhat lower with L equal to 2 instead of 3.

5.1.4. *Evaluation of the Three-Step Procedure*

An issue that we did not yet address is whether the third step in our three-step procedure is important, or whether the two-step procedure used by Vermunt (2003) performs equally well. Recall that in the third step the number of lower-level classes is reestimated, which accounts for the multilevel data structure via the higher-level classes. Table 7 reports the differences in hit rates between steps 3 and 2. As can be seen overall, the third step increases the hit rate with at most 1 percent. AIC is an

TABLE 5
Percentage of Simulation Replications in Which the Number Classes at Both Levels is Correctly Estimated

	n_k				K			H		L		p_x		p			I		
	5	10	20	50	30	100	1000	2	3	2	3	0.7	0.8	0.7	0.8	0.9	6	10	Total
BIC(N)	37	54	71	88	44	60	84	80	45	62	63	57	68	45	67	76	58	67	63
BIC(K)	42	60	79	93	54	67	85	86	52	66	71	64	74	53	74	79	64	73	69
BIC(N, K)	40	58	75	90	48	64	85	82	50	66	65	61	70	48	70	79	61	71	66
CAIC(N)	43	57	72	86	43	63	87	82	46	63	65	59	69	48	68	77	60	68	64
CAIC(K)	39	58	76	92	51	63	85	84	49	65	68	61	72	50	71	78	62	71	66
CAIC(N, K)	40	57	74	89	47	63	85	82	48	65	65	60	70	49	69	78	61	69	65
AIC	40	50	53	60	44	50	58	63	39	51	50	46	55	43	52	57	67	35	51
AIC3	47	63	80	91	55	70	87	87	54	68	73	66	75	59	74	79	68	73	70

TABLE 6
Percentage of Simulation Replications in Which the Number of Classes at Both Levels is Correctly Estimated Per Entropy Quintile

	$R^2_{entropy,high}$ quintile					$R^2_{entropy,low}$ quintile				
	1	2	3	4	5	1	2	3	4	5
BIC(N)	15	46	73	86	92	34	58	72	76	73
BIC(K)	18	54	84	92	95	42	67	78	81	76
BIC(N, K)	18	52	80	86	91	36	61	76	80	76
CAIC(N)	19	50	74	86	91	37	60	72	77	74
CAIC(K)	16	50	80	90	94	38	64	76	78	75
CAIC(N, K)	16	51	79	86	91	38	61	74	78	75
AIC	25	45	61	63	59	44	52	54	52	52
AIC3	23	60	85	91	93	50	69	78	79	75

TABLE 7
Difference in Percentage of Correct Number of Classes at Both Levels Between Step 3 and Step 2: Total and for Four Specific Conditions

		$I = 6, p = 0.7$		$I = 10, p = 0.9$	
	Total	$K = 1000$, $n_k = 5$	$K = 100$, $n_k = 50$	$K = 1000$, $n_k = 5$	$K = 100$, $n_k = 50$
BIC(N)	1	3	18	0	0
BIC(K)	1	0	0	0	0
BIC(N, K)	1	3	18	0	0
CAIC(N)	0	0	20	0	0
CAIC(K)	1	3	0	0	0
CAIC(N, K)	0	0	20	0	0
AIC	−3	0	0	3	−15
AIC3	1	2	3	0	0

exception, because with that measure step 3 does not increase and in fact decreases the hit rate.

An improvement of 1 percent is indeed small, but note step 3 can be expected to have an effect only in specific situations. That is, when the lower-level classes are poorly separated, information on the higher-level classes may help in finding the correct number of lower-level classes, provided that the higher-level classes are well separated themselves. To illustrate this issue, Table 7 presents the improvement in step 3 for four selected conditions. Note that the ($I = 6$, $p = 0.7$) and ($I = 10$,

$p = 0.9$) conditions represent poor and very good separation between lower-level classes, and ($K = 1000; n_k = 5$) and ($K = 100; n_k = 50$) poor and very good separation between higher-level classes, where the latter two are such that the overall lower-level sample size remains constant. It can indeed be seen that step 3 does not add anything with very well-separated lower-level classes. However, with poorly separated lower-level classes, step 3 is very important, especially with well-separated high-level classes.

We also compared the results of the proposed three-step procedure with the approach used by Bijmolt and colleagues (2004) in which models with all relevant combinations of lower- and higher-level classes are estimated. With all measures, the hit rates are the same for both procedures. An exception is AIC, for which the hit rate with the three-step procedure is 2 percent higher.

5.2. *Study II: Continuous Indicators*

The key difference between Study II and Study I is that the indicators are continuous variables. Another minor difference is that Study II investigated only the (more difficult) $I = 6$ condition, which is something that should be taken into account when comparing the results of the two studies. The main issue we are interested in is whether the results found in Study I generalize to the situation in which responses are continuous instead of categorical.

Table 8 presents the results on the simultaneous decision about the lower- and higher-level classes for Study II. The last column of this table provides the total for Study I for the $I = 6$ condition (which was also reported in the column $I = 6$ in Table 5). Comparison of the totals with continuous and categorical indicators shows that for most indices the hit rates are higher with continuous indicators. Exceptions are AIC and AIC3. AIC performs very poorly with continuous indicators. Closer inspection of the results showed that the problem already occurs in step 1 in which the number of lower-level classes is very likely to be overestimated. AIC3 does not perform as well as with categorical indicators, and BIC(K) is the preferred method now. It should be noted that the $I = 6$ condition is in fact the least favorable one for BIC(K), which implies that the difference between AIC3 and BIC(K) can be expected to be larger with $I = 10$.

TABLE 8
Percentage of Simulation Replications in Which the Number of Classes at Both Levels is Correctly Estimated (Study II)

	n_k				K			H		L		p_x		d				Total
	5	10	20	50	30	100	1000	2	3	2	3	0.7	0.8	0.4	0.7	1.0	Total	Study I ($I = 6$)
BIC(N)	35	54	69	86	43	60	80	80	41	61	61	55	67	40	67	76	61	58
BIC(K)	40	59	77	93	52	66	83	85	49	65	69	62	72	49	74	78	67	64
BIC(N, K)	38	57	74	87	46	63	83	81	47	65	63	59	69	42	71	78	64	61
CAIC(N)	41	57	71	85	43	62	85	82	45	62	65	58	69	43	69	78	63	60
CAIC(K)	36	58	74	92	49	64	82	83	47	64	66	59	70	45	71	78	65	62
CAIC(N, K)	38	57	73	87	46	62	83	82	45	64	63	58	69	43	70	78	64	61
AIC	23	36	37	39	31	33	37	41	26	33	34	32	35	32	33	36	34	67
AIC3	46	61	75	83	53	68	78	83	50	63	70	60	73	54	71	74	66	68

The dependence of the hit rates of the various information criteria on the design factors is the same as with categorical indicators. Design factor levels corresponding to larger sample sizes and better separated classes have the highest hit rates.

6. TWO EMPIRICAL EXAMPLES

We illustrate the three-step model selection procedure with two examples, one with categorical and one with continuous indicators.

6.1. *Job Satisfaction Measured with Categorical Indicators*

In the first application, we analyze one of the annual surveys conducted among university graduates by the AlmaLaurea consortium—specifically, the questionnaire items on job satisfaction answered by the summer 2004 graduates of the University of Florence (AlmaLaurea 2006). Information is available for 826 graduates having a job at the moment of interview and belonging to 23 study programs, where the smallest number of graduates per program is 8 and the largest is 155. The 12 dichotomous questionnaire items of interest measure the following aspects of the satisfaction with the current job: stability, correspondence with the major taken in university, competence/professionalism, prestige, cultural interests, social utility, independence, involvement in the working activity and in the decisional processes, schedule flexibility, salary, and career as well as the overall satisfaction. The aim of the multilevel LC analysis is to cluster graduates into classes based on their responses to the satisfaction items as well as to cluster programs based on the distribution of graduates across the graduate-level satisfaction classes.

Table 9 summarizes the results obtained with our three-step model-fitting procedure. In step 1 (where the hierarchical data structure is ignored), BIC(N) and CAIC(N) select a model with four lower-level classes, BIC(K) and CAIC(K) select a model with 5 classes, AIC3 selects a model with 8 classes, and AIC selects a model with 9 classes. For step 2, we estimated multilevel LC models with $L = 4$, $L = 5$, and $L = 8$ (we did not proceed with the AIC result $L = 9$). Irrespective of the value of L and the information criterion that is used, a model with 2 classes at the

TABLE 9
Selected Models in Steps 1, 2, and 3 with the University of Florence Data

Information Criterion	L, Step 1	H, Step 2	L, Step 3
BIC(K)	5	2	8
BIC(N)	4	2	4
BIC(N, K)	4	2	4
CAIC(K)	5	2	6
CAIC(N)	4	2	4
CAIC(N, K)	4	2	4
AIC3	8	2	8

program level should be preferred. In step 3, we estimated models with 2 LCs at the program level and different numbers of LCs at the graduate level. BIC(N), CAIC(N), and AIC3 select the same solution as in step 1, whereas BIC(K) and CAIC(K) select models with a larger number of lower-level classes (8 and 6, respectively). The explanation for the fact that these criteria select a larger number of lower-level classes in step 3 is that the higher-level separation is very good (ranging from .86 in the model with $L = 6$ to .89 in the model with $L = 8$). Note that BIC(K) and AIC3 come up with the same final conclusion, which is the result of the fact that their penalties are very similar: $\log 23 = 3.14$, which is very close to 3.

Of course, it is not only fit indices that are important for model selection, but also the interpretability of the obtained solutions. Because the solution with 8 LCs at the lower level is somewhat difficult to interpret, we will describe the solution with 4 lower-level and 2 higher-level classes. Lower-level class 1 contains the graduates who are satisfied with all aspects of the current job and class 4 the ones who are dissatisfied with all job aspects. The other two classes are satisfied with some and dissatisfied with other aspects: Class 2 is dissatisfied with job stability, salary, and career opportunities, and class 3 with correspondence with the major taken in university and cultural interests.

At the program level there are two classes, where class 1 is the larger of the two [$P(w_k = 1) = 0.81$]. Table 10 shows how the two classes differ in terms of their student-level class membership probabilities $P(x_{kj} = l|w_k = h)$. As can be seen, programs belonging to class 1 score much better in terms of the satisfaction of their graduates than programs belonging to class 2. Compared to the latter, the former have a much larger proportion of graduates belonging to the satisfied lower-level

TABLE 10
Distribution of Student-Level Classes Within Program-Level Classes for the $H = 2$ and $L = 4$ Model Estimated with the University of Florence Data

	h	
l	1	2
1	0.63	0.33
2	0.17	0.20
3	0.10	0.25
4	0.09	0.22

LC 1, and a much smaller proportion of students belonging to the dissatisfied lower-level LC 4. Also the proportions of graduates in the partially dissatisfied classes 2 and 3 are slightly smaller.

6.2. *Intelligence Measured with Continuous Indicators*

The second application uses the Van Peet (1992) data set that was used by Hox (2002) to illustrate multilevel factor analysis and by Vermunt (2008, 2010) to illustrate various types of multilevel mixture models. The data set contains six continuous measures that are supposed to be connected to intelligence: "word list," "cards," "matrices," "figures," "animals," and "occupations." Information is available for 269 children belonging to 49 families. The aim of the multilevel LC analysis is to cluster both children and families based on childrens' responses.

Table 11 shows the models selected in the various steps of our three-step procedure. In step 1, CAIC(N) and CAIC(N, K) select a

TABLE 11
Selected Model in Steps 1, 2, and 3 with the Intelligence Data

Information Criterion	*L*, Step 1	*H*, Step 2	*L*, Step 3
BIC(K)	4	3	4
BIC(N)	3	3	4
BIC(N, K)	3	3	4
CAIC(K)	3	3	4
CAIC(N)	2	2	2
CAIC(N, K)	2	2	2
AIC3	4	3	4

TABLE 12
Distribution of Children-Level Classes Within Family-Level Classes and Class-Specific Means for the $H = 3$ and $L = 4$ Model Estimated with the Intelligence Data

	h								
l	1	2	3	Word List	Cards	Figures	Matrices	Animals	Occupations
1	0.74	0.17	0.07	31.9	36.0	29.0	34.0	30.5	29.1
2	0.25	0.59	0.01	29.4	30.0	26.1	29.9	28.4	28.5
3	0.00	0.24	0.03	25.5	22.5	22.2	26.8	24.1	25.5
4	0.02	0.01	0.90	26.2	34.0	27.2	28.8	21.5	23.0

model with two lower-level classes; BIC(N), BIC(N, K), and CAIC(K) select a model with three lower-level classes; and AIC3 and BIC(K) select a model with four lower-level classes. In step 2, models with three higher-level classes are selected by all indices except CAIC(N) and CAIC(N, K), which select models with two higher-level classes. In step 3, the number of lower-level classes changes from three to four for BIC(N) and CAIC(K). After step 3, all three BICs, CAIC(K), and AIC3 select the model with three classes at the higher level and four classes at the lower level, which is also the model selected by estimating models with all relevant combinations of L and H (Vermunt 2008, 2010).

Table 12 reports the parameter estimates for the model with $H = 3$ and $L = 4$. The means of the six intelligence indicators are nicely ordered across child-level classes 1 to 3. These can therefore be labeled high, middle, and low. Children in class 4 show a somewhat mixed pattern: They perform better than the middle class on cards and figures, better than the low class on word list and matrices and worse than the low class on animals and occupations. The estimates of the lower-level class membership probabilities for the higher-level classes show that in family-level class 3 almost all children belong to the mixed child-level class. Children from families belonging to family-level class 1 are more likely to be in the high intelligence class and children from family-level class 2 are more often in the middle and low intelligence classes. These results show that there is a very strong family effect on the performance of children on these six intelligence subtests.

7. CONCLUSION

The purpose of the current study on multilevel LC models was twofold: to evaluate the performance of a new three-step model-fitting procedure and to investigate the performance of information criteria for simultaneously deciding about the number of lower- and higher-level classes.

As far as the performance of the three-step procedure is concerned, the simulation study provided evidence that the proposed model-fitting strategy works rather well. It is an improvement over the two-step procedure used by Vermunt (2003) when lower-level classes are poorly separated and higher-level classes well separated, which is the situation in which the multilevel data structure may help to identify the lower-level classes. In the two applications we also saw that the additional third step may matter. What can also be said is that the third step will never harm.

Furthermore, the three-step procedure performed equally well as model selection by estimating models for all relevant combinations of L and H, as used by Bijmolt and colleagues (2004) and Vermunt (2008). Since the three-step procedure is computationally less demanding and, moreover, allows the use of different measures for deciding about L and H, we think that the three-step approach is the preferred approach.

Regarding the sample size definition for BIC and CAIC, our simulation studies clearly showed that the number of groups (K) is the only appropriate sample size for deciding about the number of higher-level classes, which is in agreement with the results reported by Lukočienė and Vermunt (2010). For the decision about the number of lower-level classes, it makes less of a difference which sample size is used, but somewhat surprisingly here BIC(K) and CAIC(K) also perform slightly better than BIC(N) and CAIC(N).

Overall, AIC3 and BIC(K) turn out to be the preferred criteria for simultaneously deciding about the number of lower- and higher-level classes in models with categorical indicators. The good performance of AIC3 with categorical responses is in agreement with simulation results for standard LC models (Andrews and Currim 2003; Dias 2004; Sarstedt 2008). However, with continuous response variables, BIC(K) performs better than AIC3. AIC performs best in very specific situations—namely, with poorly separated classes and categorical indicators. These results agree with Fonseca and Cardoso's (2007) conclusions for single-level LC models based on an overview of a large number of simulation studies.

Simulation studies such as the ones reported in this paper always have certain limitations. For example, we did not investigate models with combinations of categorical and continuous indicators. It seems that in such situations BIC(K) is the preferred criterion, assuming that these models are in fact a mixture of the two types of models that were investigated.

Another limitation is that we assumed that both variables to be included in the LC analysis and the model to be used within LCs are known. More specifically, we did not consider variable selection methods for LC analysis such as the one proposed by Dean and Raftery (2009), nor did we consider tools for detecting dependencies between variables within classes as proposed by Hagenaars (1988) and others. It should, however, be noted that these more advanced tools could be used in step 1 of our three-step procedure, in which standard LC models are in fact estimated.

REFERENCES

Aitkin, Murray. 1999. "A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models." *Biometrics* 55:117–28.

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19:716–23.

AlmaLaurea. 2006. *Condizione Occupazionale dei Laureati pre e post riforma, VIII Indagine 2005*. Bologna: Consorzio Interuniversitario AlmaLaurea.

Andrews, Rick L., and Imran S. Currim. 2003. "A Comparison of Segment Retention Criteria for Finite Mixture Logit Models." *Journal of Marketing Research* 40:235–43.

Asparouhov, Tihomir, and Bengt O. Muthén. 2008. "Multilevel Mixture Models." Pp. 27–75 in *Advances in Latent Variable Mixture Models*, edited by G. R. Hancock and K. M. Samuelsen. Charlotte, NC: Information Age Publishing.

Bartholomew, David J., and Martin Knott. 1999. *Latent Variable Models and Factor Analysis*. London: Arnold.

Bassi, Francesca. 2009. "Latent Class Models for Marketing Strategies: An Application to the Italian Pharmaceutical Market." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 5:40–45.

Bezdek, J. C., Li W. Q., Y. Attikiouzel, and M. Windham. 1997. "A Geometric Approach to Cluster Validity for Normal Mixtures." *Soft Computing—A Fusion of Foundations, Methodologies and Applications* 1:166–79.

Biernacki, Christophe. 1997. "Choix de modèles en classification." PhD dissertation, University of Technology of Compiègne, Compiègne, France.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert. 2000. "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood." *IEEE Transactions on Pattern analysis and Machine Intelligence* 22:719–25.

Bijmolt, Tammo H. A., Leo J. Paas, and Jeroen K. Vermunt. 2004. "Country and Consumer Segmentation: Multi-level Latent Class Analysis of Financial Product Ownership." *International Journal of Research in Marketing* 21:323–40.

Bouwmeester, Samantha, Jeroen K. Vermunt, and Klaas Sijtsma. 2007. "Development and Individual Differences in Transitive Reasoning: A Fuzzy Trace Theory Approach." *Developmental Review* 27:41–74.

Bozdogan, Hamparsum. 1987. "Model Selection and Akaike's Information Criterion(AIC): The General Theory and Its Analytical Extensions." *Psychometrika* 52:345–70.

———. 1993. "Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix." Pp. 40–54 in *Studies in Classification, Data Analysis, and Knowledge Organization*, edited by O. Opitz, B. Lausen, and R. Klar. Heidelberg, Germany: Springer-Verlag.

———. 1994. "Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity." Pp. 69–113 in *Proceedings of the First US/Japan Conference on Frontiers in Statistical Modeling*, edited by H. Bozdogan, S. L. Sclove, and A. K. Gupta. Amsterdam: Kluwer.

Cavrinia, Giulia, Giuliano Galimberti, and Gabriele Soffritti. 2009. "Evaluating Patient Satisfaction Through Latent Class Factor Analysis." *Health and Place* 15:210–18.

Clogg, Clifford C., and Leo A. Goodman. 1984. "Latent Structure Analysis of a Set of Multidimentional Contingency Tables." *Journal of the American Statistical Association* 79:762–71.

Cutler, A., and M. Windham. 1994. "Information-Based Validity Functionals for Mixture Analysis." Pp. 149–70 in *Proceedings of the First US/Japan Conference on Frontiers in Statistical Modeling*, edited by H. Bozdogan, S. L. Sclove, and A. K. Gupta. Amsterdam: Kluwer.

Dean, Nema, and Adrian E. Raftery. 2009. "Latent Class Analysis Variable Selection." *Annals of the Institute of Statistical Mathematics* 62:11–35.

Di, Chongzhi, and Karen Bandeen-Roche. 2008. "Multilevel Latent Class Models with Dirichlet Mixing Distribution." Working Paper No. 174, Department of Biostatistics, Johns Hopkins University (http://www.bepress.com/jhubiostat/paper174).

Dias, José M. G. 2004. *Finite Mixture Models. Review, Applications, and Computer-intensive Methods*. Groningen, Netherlands: Research School Systems, Organisation and Management, University of Groningen.

Fonseca, Jaime R. S., and Margarida G. M. S. Cardoso. 2007. "Mixture-Model Cluster Analysis Using Information Theoretical Criteria." *Intelligent Data Analysis* 1:155–73.

Fraley, Chris, and Adrian E. Raftery. 1998. "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis." *The Computer Journal* 41:578–88.

Hagenaars, Jacques A. 1988. "Latent Structure Models with Direct Effects Between Indicators: Local Dependence Models." *Sociological Methods and Research* 16:379–405.

Hagenaars, Jacques A., and Allan L. McCutcheon. 2002. *Applied Latent Class Analysis Models*. Cambridge, England: Cambridge University Press.

Henry, Kimberly L., and Bengt O. Muthén. Forthcoming. "Multilevel Latent Class Analysis: An Application of Adolescent Smoking Typologies with Individual and Contextual Predictors." *Structural Equation Modeling*.

Hox, Joop. 2002. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum.

Kragelj, Boris, and Elmar Schlutter. 2007. "'Digital Divide' Reconsidered: A Country- and Individual-Level Typology of Digital Inequality in 26 European Countries. Quantitative Methods in the Social Sciences (QMSS)." Presented at a Meeting of the European Science Foundation, June 20–23, Pragvue, Czech Republic.

Lin, Ting Hsiang, and C. Mitchell Dayton. 1997. "Model Selection Information Criteria for Nonnested Latent Class Models." *Journal of Educational and Behavioral Statistics* 22:249–64.

Lukočiené, Olga, and Jeroen K. Vermunt. 2010. "Determining the Number of Components in Mixture Models for Hierarchical Data." Pp. 241–49 in *Advances in Data Analysis, Data Handling and Business Intelligence*, edited by A. Fink, L. Berthold, W. Seidel, and A. Ultsch. Berlin-Heidelberg: Springer.

Magidson, Jay, and Jeroen K. Vermunt. 2004. "Latent Class Models." Pp. 175–98 in *The Sage Handbook of Quantitative Methodology for the Social Sciences*, edited by D. Kaplan. Thousand Oaks, CA: Sage.

McLachlan, Geoffrey. 1987. "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture." *Applied Statistics* 36:318–24.

McLachlan, Geoffrey, and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Nylund, Karen L., Tihomir Asparouhov, and Bengt O. Muthén. 2007. "Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study." *Structural Equation Modeling* 14:535–69.

Palardy, Gregory, and Jeroen K. Vermunt. Forthcoming. "Multilevel Growth Mixture Models for Classifying Group-Level Observations." *Journal of Educational and Behavioral Statistics*.

Pauler, Donna K. 1998. "The Schwarz Criterion and Related Methods for Normal Linear Models." *Biometrika* 85(1):13–27.

Pirani, Elena, Silvana Schifini and Jeroen K. Vermunt. 2009. "Poverty and Social Exclusion in Europe: Differences and Similarities Across Regions." Presented at the 26th Conference of the International Union for the Scientific Study of Populations, Marrakech.

Rindskopf, David. 2006. "Heavy Alcohol Use in the 'Fighting Back' Survey Sample: Separating Individual and Community Level Influences using Multilevel Latent Class Analysis." *Journal of Drug Issues* 36:441–62.

Sarstedt, Marko. 2008. “Market Segmentation with Mixture Regression Models: Understanding Measures that Guide Model Selection.” *Journal of Targeting, Measurement, and Analysis for Marketing* 16(3):228–46.

Schwarz, Gideon E. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6:461–64.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized Latent Variables Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC.

Snijders, Tom A. B., and Roel J. Bosker. 1999. *Multilevel Analysis*. London: Sage.

Van Peet, Arie A. J. 1992. *De potentieeltheorie van intelligentie (The potentiality theory of intelligence)*. PhD dissertation, University of Amsterdam.

Vermunt, Jeroen K. 2003. “Multilevel Latent Class Models.” Pp. 213–93 in *Sociological Methodology*, vol. 33, edited by Ross. M. Stolzenberg. Boston, MA: Blackwell Publishing.

———. 2004. “An EM Algorithm for the Estimation of Parametric and Nonparametric Hierarchical Nonlinear Models.” *Statistical Neerlandica* 58:220–33.

———. 2007. “A Hierarchical Mixture Model for Clustering Three-Way Data Sets.” *Computational Statistics and Data Analysis* 51:5368–76.

———. 2008. “Latent Class and Finite Mixture Models for Multilevel Data Sets.” *Statistical Methods in Medical Research* 17:33–51.

————. 2010. “Mixture Models for Multilevel Data Sets.” Pp. 59–81 in *Handbook of Advanced Multilevel Analysis*, edited by Joop Hox and Kyle Roberts. New York: Routledge.

Vermunt, Jeroen K., and Jay Magidson. 2008. *LG-Syntax Users’ Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont, MA: Statistical Innovations.

Wedel, Michel, and Wagner A. Kamakura. 1998. *Market Segmentation: Concepts and Methodological Foundations*. Boston: Kluwer Academic Publishers.