Advance Access publication September 23, 2010

Political Analysis (2010) 18:450–469
doi:10.1093/pan/mpq025

Latent Class Modeling with Covariates: Two Improved Three-Step Approaches

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: j.k.vermunt@uvt.nl

Researchers using latent class (LC) analysis often proceed using the following three steps: (1) an LC model is built for a set of response variables, (2) subjects are assigned to LCs based on their posterior class membership probabilities, and (3) the association between the assigned class membership and external variables is investigated using simple cross-tabulations or multinomial logistic regression analysis. Bolck, Croon, and Hagenaars (2004) demonstrated that such a three-step approach underestimates the associations between covariates and class membership. They proposed resolving this problem by means of a specific correction method that involves modifying the third step. In this article, I extend the correction method of Bolck, Croon, and Hagenaars by showing that it involves maximizing a weighted log-likelihood function for clustered data. This conceptualization makes it possible to apply the method not only with categorical but also with continuous explanatory variables, to obtain correct tests using complex sampling variance estimation methods, and to implement it in standard software for logistic regression analysis. In addition, a new maximum likelihood (ML)–based correction method is proposed, which is more direct in the sense that it does not require analyzing weighted data. This new three-step ML method can be easily implemented in software for LC analysis. The reported simulation study shows that both correction methods perform very well in the sense that their parameter estimates and their SEs can be trusted, except for situations with very poorly separated classes. The main advantage of the ML method compared with the Bolck, Croon, and Hagenaars approach is that it is much more efficient and almost as efficient as one-step ML estimation.

1 Introduction

Latent class (LC) analysis (Lazarsfeld and Henry 1968; Goodman 1974a, 1974b; McCutcheon 1987; Vermunt and Magidson 2004) and related methods such as latent profile analysis (Lazarsfeld and Henry 1968) and finite mixture modeling (McLachlan and Peel 2000) are becoming increasingly popular statistical tools in a broad range of applied fields. Applications in political science research include Blaydes and Linzer (2008), Breen (2000), Edlund (2006), Feick (1989), Hill and Kriesi (2001a, 2001b), Katz and Katz (2009), Linzer (2006), McCutcheon (1985), Moors and Vermunt (2007), and Simmons (2008). These methods are used to construct a typology or clustering based on a set of observed variables; that is, to classify observational units into a—preferably small—set of LCs. In most LC analysis applications, one not only wishes to build a measurement or classification model based on a set of responses but also to relate the class membership to explanatory variables. These latter variables are referred to as covariates, predictors, external variables, independent variables, or concomitant variables. In a more explanatory study, one may wish to build a predictive or structural model for class membership, whereas in a more descriptive study the aim would be to simply profile the LCs by investigating their association with external variables.

In the LC analysis literature, two ways for dealing with covariates have been proposed: a one-step and a three-step approach. The former involves simultaneous estimation of the LC (measurement) model of interest with a logistic regression (structural) model in which the LCs are related to a set of covariates. For categorical covariates, this method was described among others by Clogg (1981), Goodman (1974b), Haberman (1979), Hagenaars (1990, 1993), and Vermunt (1997). LC models with continuous covariates were proposed by Bandeen-Roche et al. (1997), Dayton and Macready (1988), Kamakura, Wedel, and Agrawal (1994), and Yamaguchi (2000). This one-step approach, which is similar to the multiple indicator multiple cause model developed in the context of factor analysis, is implemented in the most software packages for LC analysis.

However, the one-step approach has certain disadvantages. The first is that it may sometimes be impractical, especially when the number of potential covariates is large, as will typically be the case in a more exploratory study. Each time that a covariate is added or removed not only the prediction model but also the measurement model needs to be reestimated. A second disadvantage is that it introduces additional model building problems, such as whether one should decide about the number of classes in a model with or without covariates. Third, the simultaneous approach does not fit with the logic of most applied researchers, who view introducing covariates as a step that comes after the classification model has been built. Fourth, it assumes that the classification model is built in the same stage of a study as the model used to predict the class membership, which is not necessarily the case. It can even be that the researcher who constructs the typology using an LC model is not the same as the one who uses the typology in a next stage of the study.

What is clear is that in many applications, it is more natural to use a stepwise approach and, moreover, that sometimes it is the only reasonable way to proceed. The typical stepwise approach includes:

1. An LC model is built for a set of response variables or items. This not only involves decisions on which items and how many LCs to use in the classification model but also model specification issues such as the distribution of the items within classes and the relaxation of the local independence for certain pairs of items.
2. Subjects are assigned to LCs based on their posterior class membership probabilities that can be obtained from their observed responses and the estimated parameters of the step 1 LC model. Possible classification methods are modal, random, and proportional assignment (Goodman 1974a, 1974b, 2007; McLachlan and Peel 2000; Dias and Vermunt 2008). Modal and random assignment yield what is sometimes referred to as a hard partitioning of the sample, whereas proportional assignment yields a soft or crisp partitioning.
3. A standard multinomial logistic regression model is estimated using the step 2 class assignment as the (observed) dependent variable. Rather than using a regression model, one can also simply compute two-way tables summarizing the class membership probabilities per covariate category (e.g., for males and females, for educational levels, for age groups). When combined with proportional assignment, the latter yields Magidson and Vermunt's (2001) "inactive covariates" method (see also Van der Heijden, Gilula, and Van der Ark 1999).

Bolck, Croon, and Hagenaars (2004) demonstrated that irrespective of whether one uses modal, random, or proportional assignment, three-step approaches underestimate the relationships between covariates and class membership. More specifically, they showed that the larger the amount of classification error introduced in the second step, the larger the downward bias in the parameter estimates. Based on the same derivations, Bolck, Croon, and Hagenaars (2004) and Croon (2002) developed a method for correcting the three-step approach, which I will call the BCH method. Similar approaches were proposed by Croon (2002), Lu and Thomas (2008), and Skrondal and Laake (2001) for continuous latent variables.

The BCH three-step approach proceeds as follows: (1) the data on covariates to be included in the structural model and class assignments are summarized in a multidimensional frequency table, (2) via a matrix multiplication the frequencies counts of this table are reweighted by the inverse of the matrix of classification errors, and (3) a logistic regression model is estimated using this reweighted frequency table as if it were the observed data. Problems associated with this approach are that (1) covariates need to be categorical so that the data can be summarized in a frequency table; (2) cumbersome matrix multiplications are needed in the data preparation stage and, moreover, these need to be repeated when a new set of covariates is selected, and (3) analyzing the reweighted data using a standard logistic routine yields severely downward biased SEs, and thus too liberal significance test for the logistic regression coefficients.

The aim of this article is three-fold: (1) proposing a modified BCH procedure that removes several limitations of the original BCH approach, (2) presenting an alternative more direct three-step method, and (3) reporting the results of a simulation study that show when the various three-step methods work and when they do not.

As shown in more detail in the following, the three problems associated with the BCH approach could be tackled by applying this method to individual observations rather than a table of frequency counts. It then becomes straightforward to use continuous in addition to categorical predictors and, moreover, cumbersome data preparation steps are no longer needed. In addition, the resulting weighted likelihood function maximized for parameter estimation has the form of a pseudo-likelihood similar to the one used with complex sampling designs. This suggests that correct SEs can be obtained with the linearization (sandwich) variance estimator (Skinner, Holt, and Smith 1989) or alternatively with a jackknife variance estimator (Patterson, Dayton, and Graubard 2002). As is shown in the simulation study, use of the linearization variance estimator does remove the downward bias in the SEs, which makes the BCH procedure preferable in practice. The modified BCH procedure can be implemented in standard software for logistic regression analysis that allows for (negative) sampling weights and complex sampling variance computations.

In addition, I discuss an alternative three-step method based on a logic similar to the BCH approach, namely, that in step 3 one should take into account the classification error introduced in step 2. This new three-step maximum likelihood (ML) procedure involves defining an LC model in which the step 2 class assignment serves as a single response variable with known measurement error probabilities. In this LC model, one can introduce the relevant predictors while keeping the measurement model fixed. A similar procedure was proposed by Van den Hout and Van der Heijden (2004) in the context of data collected by randomized response questions, which also yields responses with known error probabilities. The proposed three-step method can be easily implemented in software for LC analysis that allows for parameter restrictions. Besides being more elegant, the new procedure is easier to use in practice, as well as easier to extend to more complex situations, such as models with multiple latent variables constructed separately and measurement models that differ across groups. The simulation study reported below shows that this new three-step ML method is more efficient—yields smaller SEs for the covariate effects—than the BCH approach.

The remainder of this article is organized as follows. First, I describe the standard three-step and one-step approaches for LC modeling with covariates. Subsequently, I discuss the BCH method, including various modifications of this method, as well as the new three-step ML method. Then, I report the results of a simulation study comparing the performance of the various methods. Subsequently, I present an empirical application. The article ends with a summary of the main results of the current research and a discussion of possible directions for future research.

2 LC Modeling with Covariates

2.1 *The Standard Three-Step Approach*

Let us first look at the standard three-step approach, which involves (1) estimating a standard LC model without covariates, (2) assigning subjects to LCs, and (3) estimating a logistic regression model for the LCs.

2.1.1 The standard LC model

In the following, I assume that I have an LC model for a set of K categorical responses (items). The response of subject i on item k is denoted by Y_{ik}, and the full response vector by $\mathbf{Y}_i$. The discrete LC variable is denoted by X, a particular LC by t or s, and the total number of classes by T. An LC or mixture model for $P(\mathbf{Y}_i)$ can be defined as follows (Goodman 1974a, 1974b; McCutcheon 1987; Hagenaars 1990; McLachlan and Peel 2000):

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X=t)P(\mathbf{Y}_i|X=t). \tag{1}$$

Typically, categorical responses are assumed to be independent given class membership; that is,

$$P(\mathbf{Y}_i|X=t) = \prod_{k=1}^{K} P(Y_{ik}|X=t) = \prod_{k=1}^{K}\prod_{r=1}^{R_k} \theta_{ktr}^{I(Y_{ik}=r)}, \tag{2}$$

where $I(Y_{ik}=r)=1$ if subject i gives response r on item k and 0 otherwise. The parameters to be estimated are the class proportions $\pi_t = P(X=t)$ and the multinomial parameters $\theta_{ktr} = P(Y_{ik}=r|X=t)$. ML estimation of these parameters involves maximizing the following log-likelihood function:

$$\log L_{\text{STEP1}} = \sum_{i=1}^{N} \log P(\mathbf{Y}_i) = \sum_{i=1}^{N} \log\left[\sum_{t=1}^{T} \pi_t \prod_{k=1}^{K}\prod_{r=1}^{R_k} \theta_{ktr}^{I(Y_{ik}=r)}\right]. \tag{3}$$

This defines the first step of the three-step analysis.

2.1.2 Estimating class membership and classification error

In the second step, one assigns subjects to LCs on the basis of their observed responses $\mathbf{Y}_i$ and the parameter estimates from the first step. The assigned class membership of subject i is denoted by W_i. The key quantity for the class assignment is the probability of belonging to class t given the observed responses $\mathbf{Y}_i$, or the posterior class membership probability $P(X = t|\mathbf{Y}_i)$, which can be obtained by the Bayes's rule (Goodman 1974a, 1974b, 2007; McLachlan and Peel 2000; Dias and Vermunt 2008); that is,

$$P(X = t|\mathbf{Y}_i) = \frac{P(X = t)P(\mathbf{Y}_i|X = t)}{P(\mathbf{Y}_i)}. \tag{4}$$

Note that the terms appearing at the right-hand side of this equation were defined above.

The two most widely used classification rules are modal and proportional assignment. Modal assignment estimates W_i as the value of t for which $P(X = t|\mathbf{Y}_i)$ is largest; that is, it yields a hard partitioning in which individual i is treated as belonging to class t with weight $w_{it} = P(W_i = t|\mathbf{Y}_i) = 1$ if $P(X = t|\mathbf{Y}_i)$ is largest and with weight $w_{it} = 0$ otherwise. Proportional "assignment" treats subjects as belonging to LC t with probability $P(X = t|\mathbf{Y}_i)$; that is, it yields a soft (or crisp) partitioning with weights $w_{it} = P(W_i = t|\mathbf{Y}_i) = P(X = t|\mathbf{Y}_i)$. Another classification rule is random assignment that yields a hard partitioning by estimating W_i using a random draw from $P(X = t|\mathbf{Y}_i)$ (Goodman 2007). In the following, I focus on modal and proportional assignment only.

The amount of classification error can be quantified by means of the conditional probability $P(W = s|X = t)$ expressing the probability of the estimated value conditional on the true value. Using simple probability calculus, this probability can be obtained as follows:

$$\begin{aligned} P(W = s|X = t) &= \sum_{\mathbf{Y}} P(\mathbf{Y}|X = t)P(W = s|\mathbf{Y}) \\ &= \frac{\sum_{\mathbf{Y}} P\left(\mathbf{Y}\right)P\left(X = t\middle|\mathbf{Y}\right)P\left(W = s\middle|\mathbf{Y}\right)}{P(X = t)}, \end{aligned} \tag{5}$$

where the sum is over all possible response patterns. Note that $P(W = s|\mathbf{Y})$ is either 0 or 1 with modal assignment and $P(W = s|\mathbf{Y}) = P(X = s|\mathbf{Y})$ with proportional assignment. The total proportion of classification errors equals $\sum_{t=1}^{T} P(X = t) \sum_{s \neq t} P(W = s|X = t)$.

Often, it is practical to replace the sum over all possible response patterns appearing in equation (5) by a sum over all observations in the data set used to estimate the LC model of interest, which implies that $P(\mathbf{Y})$ is replaced by its empirical distribution. This yields,

$$\begin{aligned} P(W = s|X = t) &= \frac{\sum_{i=1}^{N} P(X = t|\mathbf{Y}_i)\,P(W_i = s|\mathbf{Y}_i)}{P(X = t)} \\ &= \frac{\sum_{i=1}^{N} P(X = t|\mathbf{Y}_i)\,w_{is}}{P(X = t)}. \end{aligned} \tag{6}$$

The results obtained with equations (5) and (6) can be expected to be very similar as long as the model fits the data well and the sample size is large enough. This is investigated in more detail in the simulation study reported below.

Note that the major LC analysis software packages report classification information closely related to $P(W = s|X = t)$ for modal assignment. For example, the classification table provided by Latent GOLD (Vermunt and Magidson 2005, 2008) equals $P(X = t, W = s)$ times the sample size, from which $P(W = s|X = t)$ is easily obtained by rescaling the rows to sum to 1. Mplus (Muthén and Muthén 2007) reports $P(X = t|W = s)$ as well as the number of persons assigned to each of the LCs, from which $P(W = s|X = t)$ can also be obtained. Both programs get this information using equation (6); that is, using the empirical distribution of $\mathbf{Y}$.

2.1.3 Regressing the estimated class membership on covariates

Let Z_{iq} be one of Q covariates and $\mathbf{Z}_i$ the covariate vector for subject i. The third step of the analysis involves estimating the effect of these covariates on the estimated class membership W using a multinomial logistic regression model; that is,

$$P(W = t|\mathbf{Z}_i) = \frac{\exp(\gamma_{0t} + \sum_{q=1}^{Q} \gamma_{qt} Z_{iq})}{\sum_{s=1}^{T} \exp(\gamma_{0s} + \sum_{q=1}^{Q} \gamma_{qs} Z_{iq})}. \quad (7)$$

The parameters of interest are the γ_{qt}, for $0 \leqslant q \leqslant Q$. These are obtained by maximizing the following weighted log-likelihood function:

$$\log L_{\text{STEP3}} = \sum_{i=1}^{N} \sum_{t=1}^{T} w_{it} \log P(W = t|\mathbf{Z}_i), \quad (8)$$

where $w_{it} = P(W = t|\mathbf{Y}_i)$ was defined above. Note that this involves performing a standard multinomial logistic regression analysis using an expanded data set with T records per observation and the w_{it} as weights. However, with modal assignment, there is no need to construct such an expanded data file because $w_{it} = 1$ for the assigned class and 0 otherwise.

2.2 *The One-Step ML Approach*

It is also possible to define an LC model with covariates, which makes it unnecessary to use the above three-step approach. The covariate effects are then estimated simultaneously with the parameters defining the class-specific item distributions. Models with categorical covariates were used earlier by Goodman (1974a), Clogg (1981), and Hagenaars (1990); models with continuous covariates have been developed by Dayton and Macready (1988), Bandeen-Roche et al. (1997), and Yamaguchi (2000).

When covariates are included in the LC model, one has a model for $P(\mathbf{Y}_i|\mathbf{Z}_i)$ rather than for $P(\mathbf{Y}_i)$. More specifically, the one-step (or full-information ML [FIML] estimation) approach to LC analysis with covariates involves using a model of the form

$$P(\mathbf{Y}_i|\mathbf{Z}_i) = \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i) P(\mathbf{Y}_i|X = t), \quad (9)$$

where again local independence across the $\mathbf{Y}_i$ variables may be assumed restricting $P(\mathbf{Y}_i|X = t)$ as shown in equation (2). Note that it is also assumed that $\mathbf{Y}_i$ is independent of $\mathbf{Z}_i$ conditional on X. The probability $P(X = t|\mathbf{Z}_i)$ will typically be parameterized by means of a multinomial logistic regression model:

$$P(X = t|\mathbf{Z}_i) = \frac{\exp(\gamma_{0t} + \sum_{q=1}^{Q} \gamma_{qt} z_{iq})}{\sum_{s=1}^{T} \exp(\gamma_{0t'} + \sum_{q=1}^{Q} \gamma_{qs} z_{iq})}. \quad (10)$$

FIML estimates of the γ parameters and the multinomial parameters defining $P(\mathbf{Y}_i|X = t)$ are obtained by maximizing a log-likelihood function based on $P(\mathbf{Y}_i|Z)$; that is,

$$\log L_{\mathrm{FIML}} = \sum_{i=1}^{N} \log P(\mathbf{Y}_i|\mathbf{Z}_i) = \sum_{i=1}^{N} \log \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i)P(\mathbf{Y}_i|X = t). \tag{11}$$

This is what software for LC analysis with covariates will do.

3 The BCH Approach and Some Improvements

Bolck, Croon, and Hagenaars (2004) and Croon (2002) demonstrated that the estimated γ parameters from the three-step approach are biased toward 0 and indicated how this bias can be corrected by modifying the third step of the three-step approach. The key of their contribution is the demonstration of the relationship between $P(W = s|\mathbf{Z}_i)$ and $P(X = t|\mathbf{Z}_i)$; that is, between the probability that is modeled in the third step of the three-step approach and the probability that one intends to model.

The starting point is the joint probability $P(X = t, \mathbf{Y}, W = s|\mathbf{Z}_i)$ and its decomposition:

$$P(X = t, \mathbf{Y}, W = s|\mathbf{Z}_i) = P(X = t|\mathbf{Z}_i)P(\mathbf{Y}|X = t)P(W = s|\mathbf{Y}), \tag{12}$$

which is based on the assumptions made in the LC model with covariates—$P(X = t, \mathbf{Y}|\mathbf{Z}_i) = P(X = t|\mathbf{Z}_i)P(\mathbf{Y}|X = t)$—and in the step 2 classification procedure—$P(W = s|X = t, \mathbf{Y}, \mathbf{Z}_i) = P(W = s|\mathbf{Y})$. As always, $P(W = s|\mathbf{Z}_i)$ can be obtained from $P(X = t, \mathbf{Y}, W = s|\mathbf{Z}_i)$ by summation over all LCs X and all response patterns $\mathbf{Y}$; that is:

$$\begin{aligned} P(W = s|\mathbf{Z}_i) &= \sum_{t=1}^{T} \sum_{\mathbf{Y}} P(X = t|\mathbf{Z}_i)P(\mathbf{Y}|X = t)P(W = s|\mathbf{Y}) \\ &= \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i) \sum_{\mathbf{Y}} P(\mathbf{Y}|X = t)P(W = s|\mathbf{Y}) \\ &= \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i)P(W = s|X = t). \end{aligned} \tag{13}$$

The last equation shows that $P(W = s|\mathbf{Z}_i)$ is a linear combination of $P(X = t|\mathbf{Z}_i)$, where the classification errors serve as "regression" weights. Note that $P(W = s|X = t)$ was defined in equations (5) and (6)

Bolck, Croon, and Hagenaars (2004) used the linear relationship in equation (13) for two purposes:

1. To show that the (population) log odds ratios computed using $P(W = s|\mathbf{Z}_i)$ are always smaller (closer to 0) than those obtained from $P(X = t|\mathbf{Z}_i)$, and
2. To show how to obtain $P(X = t|\mathbf{Z}_i)$ by a linear transformation of $P(W = s|\mathbf{Z}_i)$.

In order to illustrate the second point, which is of primary interest here, let $e_{is} = P(W = s|\mathbf{Z}_i)$, $a_{it} = P(X = t|\mathbf{Z}_i)$, and $d_{ts} = P(W = s|X = t)$ be elements of matrices **E**, **A**, and **D,** respectively. Equation (13) can be expressed in matrix notation as follows:

$$\mathrm{E} = \mathrm{AD}. \tag{14}$$

Using simple matrix calculus, it can be shown that **A** can be obtained as follows:

$$\mathrm{A} = \mathrm{E}\,\mathrm{D}^{-1}, \tag{15}$$

which can be solved as long as **D** is nonsingular. The latter requires that the condition $P(W = s|X = t) = P(W = s|X = t')$ for all s does not hold for any $t \neq t'$.

In order to understand how Bolck, Croon, and Hagenaars (2004) used equation (15) to modify the last step of the three-step approach, it is important to realize that they assumed all covariates are categorical, which implies the data can be summarized in a frequency table. Let $\mathbf{Z}_j^*$ denote one of the J covariate patterns, n_{js} the number of observations with covariate pattern j assigned to LC s, and **N** the frequency table with entries n_{js}. Note that **N** contains the data used to estimate **E** in the standard implementation of the third step. The correction proposed by Bolck, Croon, and Hagenaars (2004) involves using the reweighted observed frequency table $\mathbf{N}^* = \mathbf{N}\,\mathbf{D}^{-1}$ as data matrix to obtain consistent estimates of **A**. Although they do not provide the function they are maximizing for parameter estimation, they are, in fact, using a kind of pseudo-ML estimation. With $\mathbf{D}^* = \mathbf{D}^{-1}$ and d^*_{st} being an element of $\mathbf{D}^*$, the pseudo log-likelihood function that is maximized is

$$\begin{aligned}\log L_{\mathrm{BCH}} &= \sum_j \sum_{s=1}^{T} n_{jt} \sum_{t=1}^{T} d^*_{st} \log P(X = t|\mathbf{Z}_j^*)\\ &= \sum_j \sum_{t=1}^{T} n^*_{jt} \log P(X = t|\mathbf{Z}_j^*),\end{aligned} \tag{16}$$

where $n^*_{jt} = \sum_{s=1}^{T} n_{js}\, d^*_{st}$. This shows that the application of the BCH procedure requires constructing a data set with $J \cdot T$ rows where the n^*_{jt} serve as weights and subsequently performing a logistic regression analysis in the usual way. Three limitations of this approach are that it can only be used with categorical predictors, that a new data matrix should be created each time that the set of covariates is changed, and, most importantly, that the procedure does not yield correct SEs.

It can easily be seen how to solve these three problems by writing the pseudo log-likelihood in terms of individual observations rather than weighted covariate patterns. This yields

$$\begin{aligned}\log L_{\mathrm{BCH}} &= \sum_{i=1}^{N} \sum_{s=1}^{T} w_{is} \sum_{t=1}^{T} d^*_{st} \log P(X = t|\mathbf{Z}_i),\\ &= \sum_{i=1}^{N} \sum_{t=1}^{T} w^*_{it} \log P(X = t|\mathbf{Z}_i)\end{aligned} \tag{17}$$

where w_{is} was defined above, and $w^*_{it} = \sum_{s=1}^{T} w_{is}\, d^*_{st}$. Closer inspection of this log-likelihood shows that it involves creating an expanded data matrix with T records per individual with responses $t = 1, \ldots, T$ and weights w^*_{it}. The log-likelihood can then be maximized by estimating the logistic regression model of interest using this expanded data matrix. Variances can be estimated using the sandwich estimator for clustered and weighted observations, which is also used with complex samples (Skinner, Holt, and Smith 1989). This is the correct way to take into account that each individual provides T observations weighted by w^*_{it}.

An important difference with standard pseudo-likelihood estimation is that the w^*_{it} are not all positive. More specifically, w^*_{it} will typically be negative for $s \neq t$. For parameter estimation, this means that a procedure is needed that allows for negative weights. Moreover, it should be investigated whether the sandwich estimator yields the correct SEs for the parameter estimates when weights are negative. Another issue related to the estimation of the SEs is that the weights w^*_{it} are estimates (from steps 1 and 2) themselves, which is not taken into account by the sandwich estimator. However, the fact that weights are estimates is typical for situations in which complex survey estimators are applied. One of the purposes of the simulation study described below is to determine the quality of the proposed variance estimator; that is, to check whether it works with negative weights and whether ignoring the sampling fluctuation in the w^*_{it} is harmful.

It should be noted that while Bolck, Croon, and Hagenaars (2004) indicated that SEs were underestimated in their procedure, they attributed this to the fact that the sampling fluctuation in the class assignments and the **D** matrix is neglected. In fact, the primary reason for the underestimation of the SEs is that their procedure involves maximizing a weighted log-likelihood for clustered observations.

4 A Three-Step ML Method

As shown in equation (13), the key contribution of Bolck, Croon, and Hagenaars (2004) was showing how $P(W = s|\mathbf{Z}_i)$ is related to $P(X = t|\mathbf{Z}_i)$:

$$P(W = s|\mathbf{Z}_i) = \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i)P(W = s|X = t). \tag{18}$$

Closer inspection of this equation shows that it is very similar to the LC model with covariates defined in equation (9). Two differences are that W replaces the observed item responses $\mathbf{Y}_i$ and that the error probabilities $P(W = s|X = t)$ are assumed to be known (in step 3 they need not to be estimated anymore). The model described in equation (18) is, in fact, an LC model with a single indicator with known error probabilities, which is a well-known type of LC model. The same type of model can, for example, be used for the analysis of randomized response data, where the response variable is measured with known random error induced by the researcher to protect the respondent (Van den Hout and Van der Heijden 2001).

The above results suggest an alternative implementation of a corrected third step of the three-step analysis with covariates. More specifically, correct estimates of the covariate effects can be obtain by including the covariates of interest in an LC model in which the assigned class membership serves as the single (nominal) indicator and in which the step 2 $P(W = s|X = t)$ are treated as known error probabilities. This involves maximizing the following log-likelihood function:

$$\log L_{\mathrm{ML}} = \sum_{i=1}^{N} \log \sum_{t=1}^{T} P(X = t|\mathbf{Z}_i)P(W = s|X = t). \tag{19}$$

Note that this procedure yields ML estimates for not only $P(X = t|\mathbf{Z}_i)$ but also for the γ coefficients. It can be implemented in any software for LC modeling that allows defining fixed-value constraints on the model parameters.

As in the extended BCH pseudo-likelihood procedure discussed previously, SEs may be slightly underestimated because the classification error probabilities $P(W = s|X = t)$ are treated as known, whereas in fact they are obtained with the estimated parameters of the LC model without covariates. The simulation study reported below investigates how serious this problem is.

5 A Simulation Study

5.1 *Design*

A simulation study was conducted to assess the performance of various methods for estimating covariate effects and their SEs in LC analysis. The procedures that are compared are the one-step ML approach, the standard three-step approach, the BCH approach, the BCH approach with robust SEs, and the new three-step ML approach, where the latter four methods were applied with both modal and proportional assignment.

The quality of the investigated procedures can be expected to depend on two key factors: (1) the amount of measurement error and (2) the sample size. Our situation is limited to these two key factors since (1) the necessity for the correction depends on the size of the measurement error or the uncertainty about the classification from step 2 (on the rows of matrix **D**) and (2) the certainty about the estimate of the measurement error introduced in the second step depends on the sample size.

As the population model, I used a three-class LC model with six dichotomous (low/high) responses and three numeric covariates with five categories scored −2, −1, 0, 1, and 2 (all 125 covariate combinations are assumed to be equally likely to occur). Class 1 is most likely to give high response on all six items, class 3 scores low on all items, and class 2 scores high on the first three items and low on the last three. Using the first class as the reference category, the logit parameters for the covariate effects are set to 2 and 2 for Z1, −1 and 0 for Z2, and 0 and 0 for Z3, representing large, moderate, and no effect conditions. The two intercepts are such that overall the three classes are equally likely.

The classification error (or separation between the classes) was manipulated by means of the size of the response probabilities for the most likely response. The three levels I used are. 70, .80, and .90, respectively, which correspond to misclassification proportions of .31, .15, and .04, respectively.[1] These low, moderate, and high separation conditions can also be expressed using pseudo R^2 measures for nominal variables (see, e.g., Magidson 1981): a qualitative variance-based measure (Goodman and Kruskal's tau b) yields values of .33, .63, and .88, and an entropy-based measure yields values of .36, .65, and .90. I will use the latter three values to refer to the three conditions. Note that the low-separation condition in indeed very bad, the moderate condition is what can be seen as a rather typical situation in (exploratory) LC analysis, and the high condition corresponds to a strong measurement model. Although here I manipulate the size of the classification error using the response probabilities, one can also manipulate these using factors such as the number of items, the number of classes, the class sizes, and the number of item categories.

For the sample size, I used three levels: 500, 1000, and 10,000. Here, 500 is a kind of minimal sample size for LC analysis, especially in the low-separation condition, 1000 is a typical sample size in survey research, and 10,000 is a very large sample size in which sampling fluctuation can be expected to be almost negligible.

I will compare the various methods with respect to (1) bias in the estimates of the covariate effects, (2) bias in the SE estimates, and (3) relative efficiency.

5.2 *Results*

Table 1 presents the average results across the nine conditions investigated (3 separation levels × 3 sample sizes) obtained using equation (6) to estimate the classification error.

[1]Note that these classification error proportions pertain to models without covariates. Including the covariates in the model reduces the errors to .19, .10, and .03, respectively.

Table 1 Average estimate of three of the six γ parameters, their average SE, and their SD aggregated over the nine investigated conditions

	$\gamma_{21} = 2$			$\gamma_{22} = -1$			$\gamma_{32} = 0$		
Method	*Estimate*	*SE*	*SD*	*Estimate*	*SE*	*SD*	*Estimate*	*SE*	*SD*
One-step ML	2.06	0.21	0.22	–1.03	0.13	0.14	0.00	0.10	0.11
Modal standard	1.14	0.08	0.10	–0.67	0.07	0.08	0.00	0.06	0.07
Modal BCH	1.89	0.12	0.37	–0.97	0.09	0.16	0.01	0.07	0.11
Modal BCH and sandwich	1.89	0.32	0.37	–0.97	0.15	0.16	0.01	0.11	0.11
Modal ML	1.84	0.19	0.25	–0.96	0.12	0.14	0.01	0.10	0.10
Proportional standard	0.94	0.07	0.07	–0.59	0.07	0.06	0.00	0.06	0.05
Proportional BCH	1.91	0.12	0.36	–0.98	0.09	0.15	0.00	0.07	0.11
Proportional BCH and sandwich	1.91	0.31	0.36	–0.98	0.14	0.15	0.00	0.10	0.11
Proportional ML	1.86	0.24	0.23	–0.97	0.15	0.13	0.00	0.12	0.12

These are based on 100 replications per condition. Before discussing these results, I would like to mention that almost indistinguishable results were obtained with equation (5), which confirms our expectation that averaging the errors over the empirical distribution is not a problem when the model is correct. Because these results are so similar, I will focus on the results obtained with the more practical equation (6) only.

As can be seen from Table 1,[2] the standard three-step procedures based on modal and proportional assignment perform poorly; that is, these methods yield severe downward biases in the parameter estimates. Both the BCH and the ML three-step methods reduce the parameter bias substantially but still show a slight downward bias. The one-step ML parameter estimates are slightly upward biased.

Comparison of the average estimated SE with the SD of the parameter estimates across simulation replications shows that the standard BCH method yields severe downward biases in the SEs. The sandwich variance estimator very much improves the SE estimates with the BCH method, although they are still slightly (15%) underestimated. As can be seen from the much lower SDs, the new ML method is much more efficient than the BCH method and, moreover, almost as efficient as the one-step ML estimation. Its SE estimates are somewhat underestimated with modal assignment and slightly overestimated with proportional assignment.

Thus far I looked only at the results averaged over the nine conditions. However, the results turn out to vary considerably across conditions. Table 2 reports the average estimated value for the first covariate effect (with a population value of 2) as obtained with the seven estimation methods under the nine investigated conditions. It can easily be observed that the corrected three-step methods perform better with higher separation between classes and larger sample sizes. Problematic are the conditions combining the lowest separation level ($R^2_{entr} = .36$) and the two smallest sample sizes ($n = 500$ and $n = 1000$), showing that neither the BCH nor the ML three-step method performs well when separation between classes is poor, except for extremely large sample sizes ($n = 10{,}000$).

But how can this result be explained? The explanation is that, as observed among others by Galindo-Garre and Vermunt (2006), ML estimation of LC models tends to yield

[2]Note that results are reported for three of the six covariate effect parameters; that is, the parameters for class 2. The results for the other three parameters are very similar.

Table 2 Average of the estimate of γ_{21} for each of the nine conditions

	$n = 500$			$n = 1000$			$n = 10{,}000$		
Method	$R^2_{emtr} = .36$	$R^2_{emtr} = .65$	$R^2_{entr} = .90$	$R^2_{entr} = .36$	$R^2_{entr} = .65$	$R^2_{entr} = .90$	$R^2_{entr} = .36$	$R^2_{entr} = .65$	$R^2_{entr} = .90$
One-step ML	2.19	2.10	2.08	2.05	2.02	2.06	2.00	2.01	2.00
Modal standard	0.57	1.09	1.75	0.60	1.10	1.70	0.64	1.11	1.67
Modal BCH	1.24	2.08	2.10	1.50	2.07	2.05	1.96	2.01	1.99
Modal ML	1.17	1.94	2.06	1.43	1.96	2.06	1.93	2.01	1.99
Proportional standard	0.45	0.87	1.56	0.43	0.85	1.52	0.40	0.85	1.50
Proportional BCH	1.40	2.02	2.09	1.69	2.00	2.04	1.94	2.01	1.99
Proportional ML	1.25	1.97	2.06	1.52	1.96	2.06	1.95	2.01	2.00

solutions where differences between classes are larger than the true differences, which is also an explanation for the commonly occurring boundary estimates. This is especially true when classes are weakly separated and the sample size is small.[3] When differences between classes are overestimated, the amount of classification error used in both correction methods (see equations (5) and (6)) will be underestimated. To demonstrate this, Table 3 reports the true and estimated proportions of misclassification for each of the nine conditions. As can be seen, under the $R^2_{entr} = .36$ and $n = 500$ or 1000 conditions, this number is substantially underestimated, which is why the correction methods do not work well; that is, they are too optimistic and as a result the covariate effects remain downwardly biased. In such situations, FIML of the covariates effects is clearly preferred. Note that covariates yield additional information on class membership and thus increase the separation between classes. It should, however, be noted that the low-separation condition was chosen to be rather extreme. In empirical applications, often separation levels that correspond to our moderate condition ($R^2_{entr} = .65$) or somewhat higher are encountered (e.g., also in the application presented in the following).

A similar pattern as for the parameters can be seen for the estimated SEs. The correction methods do perform poorly with $R^2_{entr} = .36$ but work very well with $R^2_{entr} = .90$. In the latter condition, the sandwich SEs for the BCH method are almost unbiased and the same applies for the SEs of the ML method. Table 4 provides more details for the $R^2_{entr} = .65$ condition

Table 3 True and estimated proportion of classification errors for all nine conditions

	$R^2_{entr} = .36$	$R^2_{entr} = .65$	$R^2_{entr} = .90$
True	0.31	0.15	0.04
$n = 500$	0.22	0.14	0.04
$n = 1000$	0.26	0.15	0.04
$n = 10{,}000$	0.31	0.15	0.04

[3]It should be noted that this is also an explanation for why the covariate effects are slightly overestimated by the one-step ML approach when classes are weakly separated and the sample size is small.

Table 4 Average of the estimated SE of γ_{21} and SD of γ_{21} for the three conditions with $R^2_{entr}=.65$

	n = 500		n = 1000		n = 10,000	
Method	*SE*	*SD*	*SE*	*SD*	*SE*	*SD*
One-step ML	0.29	0.32	0.19	0.19	0.06	0.06
Modal BCH and sandwich	0.61	0.67	0.38	0.47	0.10	0.10
Modal ML	0.31	0.36	0.22	0.22	0.07	0.08
Proportional BCH and sandwich	0.45	0.47	0.30	0.35	0.09	0.09
Proportional ML	0.39	0.34	0.27	0.20	0.09	0.07

combined with each of the three sample sizes. These confirm the overall results reported in Table 1. The sandwich SEs for the BCH method are still somewhat biased downwards. The ML method is much more efficient than the BCH estimator, but its SEs may be overestimated when combined with proportional allocation, which makes significance tests for covariate effects somewhat conservative.

6 An Application: Citizenship Types

I illustrate the various methods for using covariates in LC analysis with data from the 2005 U.S. Citizenship, Involvement, and Democracy (CID) survey (Howard, Gibson, and Stolle 2005). The CID has 1001 respondents, and I selected nine response variables and three covariates. The nine response variables were used by Dalton (2006, 2008) to measure citizen norms or, more specifically, to illustrate his claim that citizenship norms are shifting from a pattern of duty-based citizenship to engaged citizenship, which in turn alters and expands the patterns of political participation in America. The citizenship norms questionnaire items in the CID were worded as follows: "To be a good citizen, how important is it for a person to be . . . [list items]. 0 is extremely unimportant and 10 is extremely important." The nine items could be grouped into four categories: items related to participation (vote in elections, be active in voluntary organizations, be active in politics), autonomy (form his or her opinion independently of others), social order (serve on a jury if called, always obey laws and regulations, for men to serve in the military when the country is at war, report a crime that he or she may have witnessed), and solidarity (support people who are worse off than themselves).

Dalton (2006, 2008) presented a varimax-rotated two-factor principal component analysis (PCA) solution for these nine response variables. The first component represented duty-based citizenship and was strongly related to "report a crime," "always obey the law," "serve in the military," and "serve on a jury." The second component represented engaged citizenship with large loadings for "form own opinion," "support worse off," "be active in politics," and "active in voluntary groups." The item "vote in elections" loaded on both dimensions. As far as the relationship with explanatory variables is concerned, Dalton (2006, 2008) stated that seniors and Republicans emphasize a duty-based definition of citizenship and that younger Americans, Democrats, and minorities stressed engaged citizenship.

I am not claiming there is something wrong with Dalton's data analysis but what is clear is that LC analysis is a suitable method to investigate the research question of interest; that is, whether there are different types of citizenship, and if so whether age, ethnicity, and political preference is related to the typology. I will use three covariates in the LC analysis to check whether similar conclusions are obtained as Dalton obtained with his PCA. These

Table 5 Parameters of four-class model estimated with the 2005 CID survey data set: class proportions and class-specific probabilities of finding the item concerned important

	Class			
	1 = both	*2 = duty-based*	*3 = engaged*	*4 = neither*
Class proportion	0.42	0.39	0.11	0.07
Report a crime	0.99	0.99	0.51	0.32
Always obey the law	1.00	0.93	0.68	0.51
Serve in the military	0.85	0.66	0.38	0.08
Serve on a jury	0.96	0.83	0.50	0.19
Vote in elections	0.98	0.80	0.68	0.11
Form own opinion	0.97	0.79	0.86	0.32
Support worse off	0.88	0.49	0.89	0.10
Be active in politics	0.75	0.07	0.43	0.00
Active in voluntary groups	0.90	0.09	0.53	0.04

are party preference (1 = Republican; 2 = Democrat; 3 = other), age (1 = younger than 50; 2 = 50 or older), and ethnicity (1 = white; 2 = nonwhite).

Although an LC analysis could have been performed on the original 11-category items, for simplicity of exposition, I will present an analysis of dichotomized items. More specifically, I combined the scores from 0 to 6 and from 7 to 10. It should be noted that the responses are rather skewed in the sense that many respondent used scores of 7 and higher, and scores lower than 3 are seldom used. On average across the nine items, 72% of the respondents gave a score of 7 or higher. I checked whether dichotomizing at 8 or 9 yielded similar results, and this turned out to be the case. Also, an LC analysis treating the original 11-point scale items as ordinal or continuous yielded very similar LCs.

A four-class model fitted the CID data well. This model was selected by Bayesian information criterion and the residuals in all two-way tables were small. Table 5 reports the parameters of the four-class model; that is, the class proportions and the class-specific probabilities of given the higher (important) response for all items. Inspection of these estimates shows that the LC solution is similar to Dalton's PCA solution in the sense that it seems to capture the two-dimensional structure in the data. Class 1 (42% of respondents) scores high on all items, class 2 (39%) scores high (higher than classes 3 and 4) on the duty-based items, class 3 (11%) scores high (higher than classes 2 and 4) on the engaged citizenship items, and class (7%) scores low on all items. Based on this, it can concluded that four citizenship types were identified: both duty-based and engaged, duty-based, engaged, and neither duty-based nor engaged.

Table 6 presents the information on the classification errors that is used by the three-step correction methods; that is, the **D** matrix with entries $P(W = s|X = t)$ and the inverse of this matrix ($\mathbf{D}^{-1}$). As can be seen, the classifications errors are somewhat larger with proportional than with modal assignment. The BCH method uses the $\mathbf{D}^{-1}$ entries as weights in an expanded data set with four records per respondent. With modal assignment, a respondent assigned to class 2 ($W = 2$) gets weights −0.0787, 1.1271, −0.0354, and −0.0130 for its records corresponding to $X = 1$, $X = 2$, $X = 3$, and $X = 4$, respectively. Note that these weights are larger than 1 for $X = W$, may be negative when $X \neq W$, and sum to 1 across values of X. For comparison with the simulation results, it is important to report that under modal assignment the total proportion of classification errors equals .11 and $R^2_{entr} = .73$.

Table 6 Matrix with classification errors **D** and its inverse $\mathbf{D}^{-1}$ for modal and proportional assignment

	D modal assignment W					**D** proportional assignment W			
X	1	2	3	4	X	1	2	3	4
1	0.9426	0.0471	0.0104	0.0000	1	0.8818	0.0826	0.0356	0.0000
2	0.0704	0.8968	0.0220	0.0108	2	0.0890	0.8334	0.0557	0.0220
3	0.1469	0.1560	0.6675	0.0296	3	0.1340	0.1949	0.6337	0.0374
4	0.0000	0.1169	0.0258	0.8573	4	0.0002	0.1228	0.0598	0.8172

	$\mathbf{D}^{-1}$ modal assignment X					$\mathbf{D}^{-1}$ proportional assignment X			
W	1	2	3	4	W	1	2	3	4
1	1.0672	–0.0536	–0.0148	0.0012	1	1.1529	–0.1019	–0.0562	0.0053
2	–0.0787	1.1271	–0.0354	–0.0130	2	–0.1097	1.2384	–0.1000	–0.0287
3	–0.2172	–0.2451	1.5115	–0.0492	3	–0.2119	–0.3499	1.6268	–0.0651
4	0.0172	–0.1463	–0.0407	1.1698	4	0.0317	–0.1605	–0.1039	1.2327

This indicates that in terms of separation between the classes, our application is between the moderate- and high-separation conditions of the simulation study.

The **D** entries are used as fixed-parameter values in the three-step ML approach. In Latent GOLD (Vermunt and Magidson 2008) this can be achieved as follows:

```
variables
     dependent W nominal;
     independent party nominal coding=1, age nominal coding=1,
        ethnicity nominal coding=2;
     latent X nominal 4 coding=1;
  equations
     X <- 1 + party + age + ethnicity;
     W <- (D~wei) 1 | X;
     D = {0.9426 0.0471 0.0104 0.0000
            0.0704 0.8968 0.0220 0.0108
            0.1469 0.1560 0.6675 0.0296
            0.0000 0.1169 0.0258 0.8573};
```

As can be seen, a regression model is defined for X, and the entries of matrix **D** are used as "cell weights." It is also possible to use the nonrescaled classification table as cell weights.

Table 7 reports the estimates for the covariate effects on the class membership and their SEs found with the investigated methods. Class 1 (class showing both forms of citizenship) serves as the baseline, and moreover, Republican, young, and nonwhite are the reference categories for party preference, age, and ethnicity, respectively. Table 8 reports the Wald tests for the covariates effects, again for all investigated methods. Note that these test the significance of all parameters corresponding to a covariate simultaneously.

Table 7 Covariate effects and their SEs obtained with the 2005 CID survey data set, where class 1 (=both) is the reference category

	Class 2 = duty-based				*Class 3 = engaged*				*Class 4 = neither*			
Method	*Dem.*	*Other*	*Old*	*White*	*Dem.*	*Other*	*Old*	*White*	*Dem.*	*Other*	*Old*	*White*
Parameter estimates												
One-step ML	0.12	0.24	–0.21	0.26	0.85	0.75	–0.54	–0.28	–0.11	0.61	–0.50	–0.70
Modal standard	0.07	0.23	–0.20	0.26	0.68	0.54	–0.48	–0.05	–0.01	0.61	–0.47	–0.59
Modal ML	0.08	0.26	–0.22	0.34	0.87	0.66	–0.59	–0.02	–0.09	0.64	–0.52	–0.69
Modal BCH	0.08	0.25	–0.22	0.35	0.87	0.67	–0.59	–0.06	–0.02	0.67	–0.52	–0.66
Proportional standard	0.11	0.21	–0.17	0.17	0.51	0.48	–0.35	–0.15	–0.04	0.53	–0.48	–0.54
Proportional ML	0.14	0.25	–0.20	0.26	0.88	0.75	–0.52	–0.14	–0.19	0.56	–0.57	–0.72
Proportional BCH	0.12	0.24	–0.20	0.26	0.89	0.79	–0.53	–0.24	–0.09	0.61	–0.58	–0.66
SEs												
One-step ML	0.20	0.21	0.17	0.20	0.37	0.39	0.32	0.31	0.40	0.36	0.33	0.32
Modal standard	0.17	0.18	0.15	0.17	0.31	0.33	0.26	0.26	0.36	0.34	0.30	0.28
Modal BCH	0.17	0.18	0.15	0.17	0.29	0.31	0.24	0.24	0.36	0.33	0.30	0.28
Modal BCH and sandwich	0.21	0.21	0.18	0.20	0.43	0.45	0.34	0.34	0.43	0.38	0.33	0.33
Modal ML	0.21	0.21	0.18	0.20	0.41	0.43	0.33	0.33	0.42	0.38	0.35	0.32
Proportional standard	0.17	0.18	0.15	0.17	0.28	0.29	0.24	0.24	0.35	0.33	0.30	0.28
Proportional BCH	0.17	0.18	0.15	0.17	0.30	0.31	0.24	0.23	0.35	0.33	0.31	0.28
Proportional BCH and sandwich	0.20	0.21	0.17	0.19	0.43	0.44	0.32	0.31	0.41	0.36	0.33	0.32
Proportional ML	0.23	0.23	0.20	0.23	0.54	0.55	0.40	0.40	0.45	0.40	0.38	0.35

Note. Dem, Democrat.

The parameter estimates in Table 7 show that the three-step methods without corrections yield estimates that are smaller than the ones of the one-step ML approach, though in this application the attenuation is not very extreme. To give an impression of the amount of attenuation, I computed the difference in estimated class membership probabilities between old and young among white Republicans. These are 0.057, −0.013, −0.021,

Table 8 Wald test for the covariate effects for the 2005 CID survey example

	Party preference			*Age*			*Ethnicity*		
Method	*Wald*	*df*	*p value*	*Wald*	*df*	*p value*	*Wald*	*df*	*p value*
One-step ML	11.15	6	0.084	5.08	3	0.166	9.40	3	0.024
Modal standard	10.67	6	0.099	5.50	3	0.138	9.69	3	0.021
Modal BCH	16.65	6	0.011	8.11	3	0.044	14.01	3	0.003
Modal BCH and sandwich	10.06	6	0.122	5.59	3	0.133	8.81	3	0.032
Modal ML	10.48	6	0.106	5.43	3	0.143	9.64	3	0.022
Proportional standard	8.04	6	0.235	4.38	3	0.224	7.07	3	0.070
Proportional BCH	15.77	6	0.015	7.46	3	0.059	12.72	3	0.005
Proportional BCH and sandwich	9.89	6	0.129	5.71	3	0.126	8.49	3	0.037
Proportional ML	7.58	6	0.270	4.25	3	0.236	7.01	3	0.072

and −0.023 for the standard three-step proportional approach, and 0.081, −0.014, −0.035, and −0.033 for the one-step ML approach.

The parameter estimates in Table 7 show that the three-step approaches with corrections yield estimates that are close to those obtained with one-step ML estimation. Overall, it seems that proportional assignment is closer to one-step ML than modal assignment. These results are in agreement with what was found in the simulation study.

The SE estimates are also in agreement with what could be expected based on the simulation results. The standard and BCH three-step methods yield SE estimates that are too small (smaller than of the one-step approach). The sandwich SE for the BCH method and the SE of the three-step modal ML approach are close to the ones of the one-step approach. The three-step proportional ML approach yields somewhat larger SEs.

The Wald tests show that only the effect of ethnicity is significant. The BCH methods with sandwich variance estimates and the modal ML approach yield *p* values that are close to the ones of the one-step approach. The proportional ML approach yields somewhat larger *p* values and is thus somewhat too conservative. It can also be seen that the BCH three-step methods without corrected variances yield *p* values that are much too small, which in this application would lead to wrong conclusions regarding the statistical significance of the party preference and age effects.

Having a closer look at the estimated covariates effects, as well as the ratio between the parameter estimates and their SEs, shows that compared with Republicans, Democrats are more likely to be in class 3 instead of 1, and others are more likely to be in classes 3 and 4 instead of 1. Moreover, compared with the young, the old are less likely to be in classes 3 and 4 instead of class 1, and compared with nonwhites, whites are less likely to be in class 4 instead of class 1.

7 Discussion

This article proposed two improvements of the three-step method of Bolck, Croon, and Hagenaars (2004). First, it was demonstrated how it can be used with nongrouped data, which makes it possible to use the method also with continuous explanatory variables. Second, because parameter estimation involves maximizing a weighted log-likelihood for clustered data, it was proposed to estimate the SEs using complex sampling methods. In addition, a new ML-based correction method was proposed, which is based on the same logic but which is more direct in the sense that it does not require analyzing weighted data. This three-step ML method can be easily implemented in software for LC analysis.

The reported simulation study showed that both correction methods perform very well in the sense that their parameter estimates and their SEs can be trusted, except for situations with very poorly separated classes. Before applying the correction methods, it is therefore important to check whether class separation is not too low. The main advantage of the ML method compared with the BCH approach is that it is much more efficient and almost as efficient as the one-step ML approach. A minor disadvantage of the new method is that software for LC analysis is needed for step 3, whereas with the BCH approach, standard software for multinomial logistic regression with complex sampling features (and allowing for negative weights) suffices. However, given that step 1 also requires LC analysis software, this does not seem to be a big issue.

Whereas in this article I focused on simple LC models for discrete responses, the two correction methods can also be applied with other types of mixture models, for example, with mixture models for continuous variables, factor mixture models, and mixture growth models. These are all models in which it may be attractive to introduce covariates in a next step after the mixture model itself was constructed. It can be expected that similar types of conditions will determine the performance of the three-step methods, but of course this needs to investigated.

Other possible applications of the proposed three-step methods are in LC analysis for longitudinal or multilevel data; that is, as an alternative to one-step approaches such as latent Markov modeling (Van de Pol and Langeheine 1990; Collins and Wugalter 1992; Vermunt, Langeheine, and Böckenholt 1999) and multilevel LC modeling (Vermunt 2003, 2008). An LC model could first be built without taking into account the longitudinal or multilevel data structure, and the classifications with known errors could subsequently be used in step 3. In a multilevel context, the model estimated in step 3 could have the form of a random-effect logistic regression model, which could serve as an alternative to the (one-step) model proposed by Vermunt (2005).

Another issue that deserves further investigation is the effect of violations of the assumptions underlying the correction methods as well as one-step LC models with covariates. The most important of these is the assumption that covariates have no direct effects on the responses after controlling for a person's class membership (Hagenaars 1990, 1993). It could very well be that three-step methods are more robust for violations of this assumption than one-step methods.

Recently, Bayesian estimation procedures for LC models have been proposed (Garrett and Zeger 2000; Garrett, Eaton, and Zeger 2002; Chung, Flaherty, and Schafer 2006). It would be worthwhile investigating how to apply the three-step methods proposed in this article with Bayesian estimation. For example, in step 3 one could estimate the covariate effects using the random class assignments and the corresponding measurement error estimates from a Markov chain Monte Carlo sampler. This could be repeated several times, yielding a procedure similar to multiple imputation (Rubin 1987; Schafer 1997). Such a procedure would make it possible to take into account the uncertainty about the class assignments and the classification errors. It may be that such a Bayesian three-step procedure performs better than the three-step procedures discussed here, especially with small sample sizes and badly separated classes.

References

Bandeen-Roche, Karen, Diana L. Miglioretti, Scott L. Zeger, and Paul J. Rathouz. 1997. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* 92:1375–86.

Blaydes, Lisa, and Drew A. Linzer. 2008. The political economy of women's support for fundamentalist Islam. *World Politics* 60:579–609.

Bolck, Annabel, Marcel A. Croon, and Jacques A. Hagenaars. 2004. Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis* 12:3–27.

Breen, Richard. 2000. Why is support for extreme parties underestimated by surveys? A latent class analysis. *British Journal of Political Science* 30:375–82.

Chung, Hwan, Brian P. Flaherty, and Joseph L. Schafer. 2006. Latent class logistic regression: Application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society Series A—Statistics in Society* 169:723–43.

Clogg, Clifford C. 1981. New developments in latent structure analysis. In *Factor analysis and measurement in sociological research*, ed. D. J. Jackson and E. F. Borgotta, 215–46. Beverly Hills, CA: Sage.

Collins, Linda M., and Stuart E. Wugalter. 1992. Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research* 27:131–57.

Croon, Marcel A. 2002. Using predicted latent scores in general latent structure models. In *Latent variable and latent structure models*, ed. George A. Marcoulides and Irini Moustaki, 195–224. Mahwah, NJ: Lawrence Erlbaum.

Dalton, Russell J. 2006. The two faces of citizenship. *Democracy & Society* 3:21–3.

———. 2008. Citizenship norms and the expansion of political participation. *Political Studies* 56:76–98.

Dayton, C. Mitchell, and Geoffrey B. Macready. 1988. Concomitant-variable latent-class models. *Journal of the American Statistical Association* 83:173–8.

Dias, José G., and Jeroen K. Vermunt. 2008. A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics* 23:643–59.

Edlund, Jonas. 2006. Trust in the capability of the welfare state and general welfare state support: Sweden 1997–2002. *Acta Sociologica* 49:395–417.

Feick, Lawrence F. 1989. Latent class analysis of survey questions that include don't know responses. *Public Opinion Quarterly* 53:525–47.

Galindo-Garre, Francisca, and Jeroen K. Vermunt. 2006. Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika* 33:43–59.

Garrett, Elisabeth S., and Scott L. Zeger. 2000. Latent class model diagnosis. *Biometrics* 56:1055–67.

Garrett, Elisabeth S., William W. Eaton, and Scott L. Zeger. 2002. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: A latent class model approach. *Statistics in Medicine* 21: 1289–307.

Goodman, Leo A. 1974a. The analysis of systems of qualitative variables when some of the variables are unobservable: Part I—A modified latent structure approach. *American Journal of Sociology* 79:1179–259.

———. 1974b. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215–31.

———. 2007. On the assignment of individuals to classes. *Sociological Methodology* 37:1–22.

Haberman, Shelby J. 1979. *Analysis of qualitative data, Vol. 2: New developments*. New York: Academic Press.

Hagenaars, Jacques A. 1990. *Categorical longitudinal data—Loglinear analysis of panel, trend and cohort data*. Newbury Park, CA: Sage.

———. 1993. *Loglinear models with latent variables*. Newbury Park, CA: Sage.

Hill, Jennifer L., and Hanspeter Kriesi. 2001a. Classification by opinion-changing behavior: A mixture model approach. *Political Analysis* 9:301–24.

———. 2001b. An extension and test of converse's 'black-and-white' model of response stability. *American Political Science Review* 95:397–413.

Howard, Marc M., James L. Gibson, and Dietlind Stolle. 2005. *The U.S. Citizenship, Involvement, Democracy survey.* Center for Democracy and Civil Society, Georgetown University.

Kamakura, Wagner A., Michel Wedel, and Jagdish Agrawal. 1994. Concomitant variable latent class models for the external analysis of choice data. *International Journal of Marketing Research* 11:451–64.

Katz, Jonathan N., and Gabriel Katz. 2009. Reassessing the link between voter heterogeneity and political accountability: A latent class regression model of economic voting. Paper presented at the 26th Annual Society for Political Methodology Summer Conference, July 23–25, 2009, Yale University.

Lazarsfeld, Paul F., and Neil W. Henry. 1968. *Latent structure analysis*. Boston, MA: Houghton Mill.

Linzer, Drew A. 2006. A comparative analysis of ideological constraint using latent class models. Paper presented at the annual meeting of the Midwest Political Science Association, Palmer House Hilton, Chicago, IL, April 20, 2006.

Lu, Irene R. R., and D. Roland Thomas. 2008. Avoiding and correcting bias in score-based latent variable regression with discrete manifest items. *Structural Equation Modeling* 15:462–90.

Magidson, Jay. 1981. Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research* 10:177–94.

Magidson, Jay, and Jeroen K. Vermunt. 2001. Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology* 31:223–64.

McCutcheon, Allan L. 1985. A latent class analysis of tolerance for nonconformity in the American public. *Public Opinion Quarterly* 49:474–88.

———. 1987. *Latent class analysis*. Newbury Park, CA: Sage.

McLachlan, Geoffrey J., and David Peel. 2000. *Finite mixture models*. New York: Wiley.

Moors, Guy, and Jeroen K. Vermunt. 2007. Heterogeneity in postmaterialist value priorities. Evidence from a latent class discrete choice approach. *European Sociological Review* 23:631–48.

Muthén, Linda K., and Bengt O. Muthén, L. 2004. *Mplus 3.0: User's manual*. Los Angeles, CA: Muthén and Muthén.

Patterson, Blossom H., C. Mitchell Dayton, and Barry I. Graubard. 2002. Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association* 97:721–8.

Rubin, Donald B. 1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, Joseph L. 1997. *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Skinner, Chris, J., Tim Holt, and T. M. Fred Smith. 1989. *Analysis of complex surveys*. New York: Wiley.

Simmons, Solon. 2008. Ascriptive justice: The prevalence, distribution, and consequences of political correctness in the academy. *Forum* 6:8.

Skrondal, Anders, and Petter Laake. 2001. Regression among factor scores. *Psychometrika* 88:563–76.
Van den Hout, Ardo, and Peter G. M. Van der Heijden. 2004. The analysis of multivariate misspecified data, with special attention to randomized response data. *Sociological Methods and Research* 32:310–36.
Van de Pol, Frank, and Rolf Langeheine. 1990. Mixed Markov latent class models. *Sociological Methodology* 20:213–47.
Van der, Heijden, Peter G. M. Zvi Gilula, and L. Andries Van der Ark. 1999. An extended study into the relationship between correspondence analysis and latent class analysis. *Sociological Methodology* 29:147–86.
Vermunt, Jeroen K. 1997. *Log-linear models for event histories. Advanced quantitative techniques in the social sciences series*. Thousand Oaks, CA: Sage.
———. 2003. Multilevel latent class models. *Sociological Methodology* 33:213–39.
———. 2005. Mixed-effects logistic regression models for indirectly observed outcome variables. *Multivariate Behavioral Research* 40:281–301.
———. 2008. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* 17:33–51.
Vermunt, Jeroen K., Rolf Langeheine, and Ulf Böckenholt. 1999. Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics* 24:178–205.
Vermunt, Jeroen K., and Jay Magidson. 2004. Latent class analysis. In *The Sage encyclopedia of social science research methods*, ed. Michael Lewis-Beck, Alan Bryman, and Tim F. Liao, 549–53. Newbury Park, CA: Sage.
———. 2005. *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations.
———. 2008. *LG-Syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations.
Yamaguchi, Kazuo. 2000. Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among Japanese women. *American Journal of Sociology* 105:1702–40.