PSYCHOMETRIKA—VOL. 88, NO. 4, 1144–1170
DECEMBER 2023
https://doi.org/10.1007/s11336-023-09929-2

A TWO-STEP ESTIMATOR FOR MULTILEVEL LATENT CLASS ANALYSIS WITH COVARIATES

ROBERTO DI MARI
UNIVERSITY OF CATANIA

ZSUZSA BAKK
LEIDEN UNIVERSITY

JENNIFER OSER
BEN-GURION UNIVERSITY

JOUNI KUHA
LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

We propose a two-step estimator for multilevel latent class analysis (LCA) with covariates. The measurement model for observed items is estimated in its first step, and in the second step covariates are added in the model, keeping the measurement model parameters fixed. We discuss model identification, and derive an Expectation Maximization algorithm for efficient implementation of the estimator. By means of an extensive simulation study we show that (1) this approach performs similarly to existing stepwise estimators for multilevel LCA but with much reduced computing time, and (2) it yields approximately unbiased parameter estimates with a negligible loss of efficiency compared to the one-step estimator. The proposal is illustrated with a cross-national analysis of predictors of citizenship norms.

Key words: multilevel latent class analysis, covariates, stepwise estimators, pseudo ML.

Latent class analysis (LCA) is used to create a clustering of units based on a set of observed variables, expressed in terms of an underlying unobserved classification. When it is applied to hierarchical (multilevel) data where lower-level units are nested in higher-level ones, the basic latent class model can be extended to account for this data structure. This can be seen as a random coefficients multinomial logistic model (see, for instance (Agresti et al., 2000)) for an unobserved categorical variable that is measured by several observed indicators, with a higher-level latent class variable in the role of a categorical random effect (Vermunt, 2003) . Multilevel LCA has become more popular in the social sciences in recent years, for example in educational sciences (Fagginger Auer et al., 2016; Grilli et al., 2022, 2016; Grilli & Rampichini, 2011; Mutz & Daniel, 2013) , economics (Paccagnella & Varriale, 2013) , epidemiology (Tomczyk et al., 2015; Rindskopf, 2006; Zhang et al., 2012; Horn et al., 2008) , sociology (Da Costa & Dias, 2015; Morselli & Glaeser, 2018) , and political science (Ruelens & Nicaise, 2020) . In most of these examples, the multilevel LCA model includes also covariates that are used as predictors of the clustering, and substantive research questions often focus on the coefficients of the covariates.

In estimation of models with covariates, for single-level LCA the current mainstream recommendation is to use *stepwise* methods that separate the estimation of the *measurement model* for the observed indicators from the estimation of the *structural model* for the latent variables given

Correspondence should be made to Roberto Di Mari, Department of Economics and Business, University of Catania, Corso Italia 55, 95128 Catania, Italy. Email: roberto.dimari@unict.it

© 2023 The Author(s)

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

the covariates (see, e.g., (Bakk & Kuha, 2018; Di Mari et al., 2020; Di Mari & Maruotti, 2022; Vermunt, 2010)). This is practically convenient because when changes of covariates are made, only the structural model rather than the full model needs to be re-estimated. Different structural models can be considered even by different researchers at different times. Stepwise estimation can also avoid biases which can arise when all the parameters are instead estimated together in a simultaneous (*one-step*) approach to estimation. In such cases, misspecifications in one part of the model can cause bias also in the parameter estimates in other parts (Bakk & Kuha, 2018) .

In multilevel LCA, the one-step approach is particularly cumbersome because of increased estimation time, especially with multiple covariates possibly defined at different levels. In that context, there is still need for further research on bias-adjusted efficient stepwise estimators. Recently Bakk et al. (2022) and Di Mari et al. (2022) proposed a "two-stage" estimator for this purpose. The parameters of the measurement model are estimated in its first stage, without including the covariates. This is further broken down into three steps. In the first of them, initial estimates of the measurement model are obtained from a single-level LC model, ignoring the multilevel structure. The latent class probabilities of the multilevel LC model are then estimated, keeping the measurement parameters from the first step fixed. Third, to stabilize the estimated measurement model and to account for possible interaction effects, the multilevel model is estimated again, now keeping the latent class parameters fixed. The estimated measurement parameters from this last step of the first stage are then held fixed in the second stage, where the model for the latent classes given covariates is estimated.

This method has been shown to greatly simplify model construction and interpretation compared to the one-step estimator, with almost identical results if model assumptions are not violated, and with enhanced algorithmic stability and improved speed of convergence. In addition, the two-stage estimator exhibits an increased degree of robustness compared to the simultaneous approach in the presence of measurement noninvariance (Bakk et al., 2022) .

A difficulty in this two-stage technique is deriving an asymptotic covariance matrix that takes into account the multi-step procedure. Conditioning on the first-stage estimates as if they were known, even though they are estimates with a sampling distribution, introduces a downward bias in the standard errors, a phenomenon that is well known also in the context of stepwise structural equation models (Skrondal & Kuha, 2012; Oberski and Satorra, 2013) . For two-step single level LCA, the standard errors can be corrected in a straightforward way (Bakk & Kuha, 2018) , but this is more difficult for two-stage LCA due to conditioning on multiple steps.

The two-stage approach is still in some ways more involved than it needs to be. In this paper we show that it is possible to simplify it into a more straightforward *two-step estimator*, still retaining its good performance but with a further reduced computation time. This approach is closely motivated by two-step estimation as it is used for single-level LCA. In the first step, the full multilevel measurement model is estimated in one go, but without covariates. In the second step, covariates are included in the model, keeping the measurement model parameters fixed at their estimates from the first step.

With such a two-step estimator, we contribute to the existing literature in several ways: (1) we establish model identification for the multilevel LC model under standard assumptions, as foundation for correct measurement model estimation; (2) we derive a step-by-step EM algorithm with closed-form formulas to handle the computation of the two-step estimator; and (3) we derive the correct asymptotic variance-covariance matrix of the second step estimator of the structural model, drawing on the theory of pseudo maximum likelihood estimation (Gong and Samaniego, 1981) .

We evaluate the finite sample properties of our proposal by means of an extensive simulation study. Cross-national data on citizenship norms from the International Association for the Evaluation of Educational Achievement survey are analyzed to illustrate the proposal, and possible extensions are discussed in the conclusions.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

1. The Multilevel Latent Class Model with Covariates

Let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijH})'$ be a vector of observed responses, where Y_{ijh} denotes the response of individual $i = 1, \ldots, n_j$ in group $j = 1, \ldots, J$ on the h-th categorical indicator variable ("item"), with $h = 1, \ldots, H$. The data have a hierarchical (multilevel) structure where the individuals are nested within the groups. In the following, we will also refer to individuals as the "low-level units", and groups as the "high-level units". Let $\mathbf{Y}_j = (\mathbf{Y}_{1j}, \ldots, \mathbf{Y}_{n_j j})'$ denote the set of responses for all the low-level units belonging to high-level unit j, with $\mathbf{Y}_j$ for different j taken to be independent of each other. For simplicity of exposition, we focus below on the case where the items Y_{ijh} are dichotomous, but the idea and methods of two-step estimation proposed here apply in a straightforward way also for polytomous items.

Let W_j be a categorical latent variable (i.e. a *latent class* (LC) variable) defined at the high level, with possible values $m = 1, \ldots, M$ and probabilities $P(W_j = m) = \omega_m > 0$, and let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)'$. Given a realization of W_j, let X_{ij} be a categorical latent variable defined at the low level, with possible values $t = 1, \ldots, T$, and conditional probabilities $P(X_{ij} = t|W_j = m) = \pi_{t|m} > 0$. We collect all the $\pi_{t|m}$ in the $M \times T$ matrix Π. The X_{ij} for the same j are taken to be conditionally independent given W_j, so that

$$P(X_{1j}, \ldots, X_{n_j j}) = \sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} P(X_{ij}|W_j = m).$$

This shows that the high-level latent class W_j serves as a categorical random effect which accounts for associations between the low-level latent classes X_{ij} for different low-level units i within the same high-level unit j.

The items $\mathbf{Y}_j$ are treated as observed indicators of the latent classes. A multilevel latent class model specifies the probability of observing a particular response configuration for a high-level unit j as

$$\begin{aligned} P(\mathbf{Y}_j) &= \sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m)\, P(\mathbf{Y}_{ij}|X_{ij} = t, W_j = m) \\ &= \sum_{m=1}^{M} \omega_m \prod_{i=1}^{n_j} \sum_{t=1}^{T} \pi_{t|m} \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t, W_j = m), \end{aligned} \tag{1}$$

where $P(Y_{ijh}|X_{ij} = t, W_j = m)$ denotes the conditional probability mass function of the h-th item, given the latent class variables X_{ij} and W_j. The second line in this further assumes that the responses for Y_{ijh} for different items h are conditionally independent given (X_{ij}, W_j), a standard assumption which we make throughout.

Model (1) is a general formulation which is equal to an unrestricted multi-group latent class model. Most applications, however, use a more restricted version which assumes that the item response probabilities do not depend directly on the high-level latent class W_j ((Vermunt, 2003; Lukociene et al., 2010) ; this model is represented in Fig. 1, if we omit the covariates Z_{ij} which will be introduced below). We will also make this assumption throughout this paper. Model (1) is also similar to the multilevel item response model of Gnaldi et al. (2016), but with categorical latent variables at both levels. The response probabilities are then given by

$$P(\mathbf{Y}_j) = \sum_{m=1}^{M} \omega_m \prod_{i=1}^{n_j} \sum_{t=1}^{T} \pi_{t|m} \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t). \tag{2}$$

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

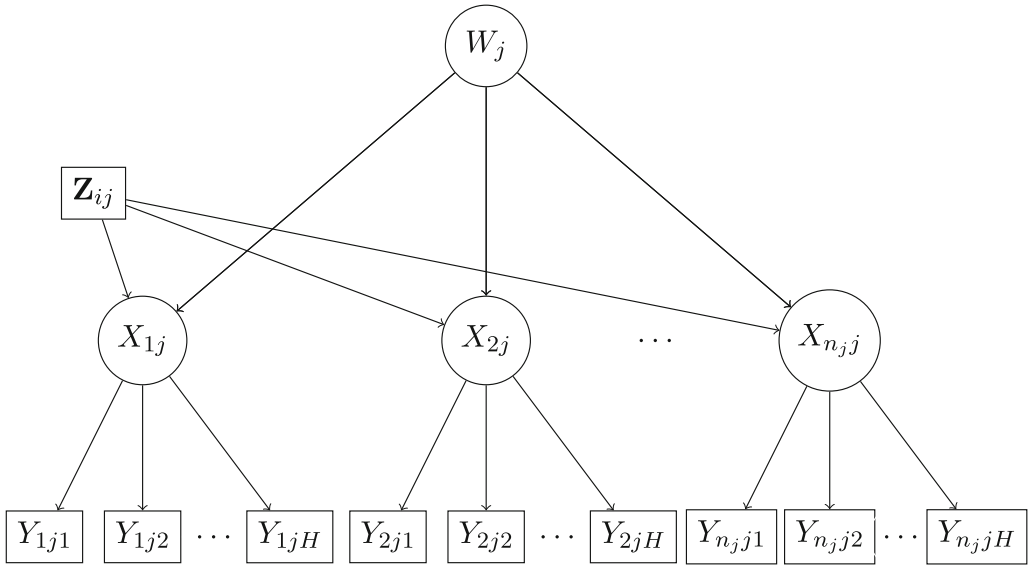


FIGURE 1.
Graphical representation of a multilevel latent class model which includes a low-level latent class variable X_{ij} nested in a high-level latent class variable W_j, and covariates Z_{ij} for X_{ij}. Here the response probabilities for items Y_{ijh} depend directly only on X_{ij}.

Therefore, within each high-level latent class W_j, the model for the items has the form of a standard (single-level) LC model with X_{ij} as the latent class (McCutcheon, 1987; Goodman, 1974; Hagenaars, 1990) . When the items Y_{ijh} are binary with values 0 and 1, we denote $P(Y_{ijh} = 1|X_{ij} = t) = \phi_{h|t}$, so that $P(Y_{ijh} = y_{ijh}|X_{ij} = t) = \phi_{h|t}^{y_{ijh}}(1 - \phi_{h|t})^{1-y_{ijh}}$, and denote by $\mathbf{\Phi}$ the $H \times T$ matrix of all the $\phi_{h|t}$.

It can be shown that the model is identified (in a generic sense, see Allman et al. 2009), under a standard set of assumptions:

Proposition 1.1. (Identification) *Suppose that the following conditions hold: (A.1)* $\phi_{h|t} \neq \phi_{h|s}$ *for all* $h = 1, \ldots, H$ *and for* $t \neq s$*; and (A.2) the* $M \times T$ *matrix* Π *has rank* M*. Then the multilevel LC model (2) is identified when* $M \leq T$ *and* $n_j \geq 3$*, for all* $j = 1, \ldots, J$*.*

The proof of Proposition 1.1 follows the same lines as in Gassiat et al. (2016), who proved identification of finite state space nonparametric hidden Markov models, and applies the results of Theorem 9 of Allman et al. (2009). The fact that all $\phi_{h|t}$ are distinct is sufficient for linear independence of the Bernoulli random variables. For $n_j = 3$, using the assumption of conditional independence of low-level units given high-level class W_j, the distribution of $(\mathbf{Y}_{1j}, \mathbf{Y}_{2j}, \mathbf{Y}_{3j})$ factorizes as the product of three terms $\mu_{ij|m} = \sum_t \pi_{t|m} P(\mathbf{Y}_{ij}|X_{ij} = t)$ for $i = 1, 2, 3$. Assumption (A.2) ensures that $\mu_{1j|m}$, $\mu_{2j|m}$ and $\mu_{3j|m}$ are linearly independent. Thus Theorem 9 of Allman et al. (2009) applies.

We make three ancillary comments on Proposition 1.1. First, for the unrestricted multilevel LC model (1), if an assumption analogous to (A.1) holds—i.e. if all success probabilities of the Bernoulli random variables are distinct—we can relax (A.2) and prove identification using Allman et al. (2009)'s Theorem 9 (in the related context of mixture of finite mixtures with Gaussian components, a similar argument is used by (Di Zio et al., 2007)). Second, for longitudinal and multilevel data, generic identification of the measurement model does not require any condition on the number of items, provided that conditions (A.1) and (A.2) are satisfied. Third, although

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

we have discussed identification specifically for binary items and Bernoulli conditional distributions, the identification result extends also to polytomous items if we can assume, analogously to (A.1), that all conditional category-class response probabilities are distinct. This guarantees linear independence of the corresponding multinomial random variables.

Covariates can be included in the multilevel LC model to predict latent class membership in both the low and high-level classes. Let $\mathbf{Z}_{ij} = (1, \mathbf{Z}'_{1j}, \mathbf{Z}'_{2ij})'$ be a vector of K covariates, which can include high-level ($\mathbf{Z}_{1j}$) and low-level ($\mathbf{Z}_{2ij}$) variables. For X_{ij} we can consider the multinomial logistic model

$$P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) = \frac{\exp(\boldsymbol{\gamma}'_{tm}\mathbf{Z}_{ij})}{1 + \sum_{s=2}^{T} \exp(\boldsymbol{\gamma}'_{tm}\mathbf{Z}_{ij})}, \tag{3}$$

where $\boldsymbol{\gamma}_{tm}$ is a K-vector of regression coefficients for each $t = 2, \ldots, T$ and $m = 1, \ldots, M$. When only the intercept term is included, so that $\mathbf{Z}_{ij} = 1$, then $\boldsymbol{\gamma}_{tm} = \log(\pi_{t|m}/\pi_{1|m})$ in the notation of the model without covariates above. We denote by $\boldsymbol{\Gamma}$ the $(T-1)M \times K$ matrix of all the parameters in the $\boldsymbol{\gamma}_{tm}$ vectors.

A model for W_j can be specified similarly, now using only high-level covariates $\mathbf{Z}^*_j = (1, \mathbf{Z}'_{1j})'$, as

$$P(W_j = m|, \mathbf{Z}^*_j) = \frac{\exp(\boldsymbol{\alpha}'_m\mathbf{Z}^*_j)}{1 + \sum_{l=2}^{M} \exp(\boldsymbol{\alpha}'_m\mathbf{Z}^*_j)}, \tag{4}$$

where $\boldsymbol{\alpha}_m$ for $m = 2, \ldots, M$, are regression coefficients. Although this too is straightforward, for ease of exposition and simplicity of notation we will below not consider models with covariates for W_j, but present the two-step estimator only for the case where $\mathbf{Z}^*_j = 1$ and thus $\boldsymbol{\alpha}_m = \log(\omega_m/\omega_1)$. The focus of interest is then on the model for the low-level (individual-level) latent class X_{ij}, and the high-level (group-level) latent class W_j serves primarily as a random effect which accounts for intra-group associations between X_{ij}. We further assume that the observed items $\mathbf{Y}_j$ are conditionally independent of the covariates $\mathbf{Z}_{ij}$ given the latent class variables X_{ij}. This means that the measurement of X_{ij} by $\mathbf{Y}_{ij}$ is taken to be invariant with respect to the covariates. With these assumptions, and denoting $\mathbf{Z}_j = (\mathbf{Z}'_{1j}, \ldots, \mathbf{Z}'_{n_j j})'$, the model that we will consider is finally of the form

$$P(\mathbf{Y}_j|\mathbf{Z}_j) = \sum_{m=1}^{M} \omega_m \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t); \tag{5}$$

see also a graphical representation of the model in Fig. 1. This model is identified when the corresponding model without covariates is identified, as long as the design matrix of all the $\mathbf{Z}_{ij}$s has full column rank (for an analogous condition for identifiability in the context of single-level latent class models with covariates, see Huang and Bandeen-Roche 2004 and Ouyang and Xu 2022).

2. Previous Methods of Estimation

We denote the parameters of the model in (5) as $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)'$ where $\boldsymbol{\theta}_1 = \text{vec}(\boldsymbol{\Phi})$ are the parameters of the measurement model for the items $\mathbf{Y}_j$ and $\boldsymbol{\theta}_2 = (\text{vec}(\boldsymbol{\Gamma})', \boldsymbol{\omega}')'$ the parameters

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

of the structural model the latent class variables (X_{ij}, W_j) given the covariates $\mathbf{Z}_{ij}$. Maximum likelihood estimates of these parameters can be obtained by maximizing the log likelihood $\ell(\boldsymbol{\theta}) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j|\mathbf{Z}_j)$ with respect to all the parameters together. This is the simultaneous or *one-step method* of estimation for the model. It has serious disadvantages, however. The full model needs to be re-estimated whenever the covariates in the structural model are changed, which can be computationally demanding because of the complexity of such multilevel models. Further, because all the parameters are estimated together, misspecification in one part of the model may destabilize also parameters in other parts of the model (Vermunt, 2010; Asparouhov and Muthén, 2014) .

Because of the complexity of the one-step approach, in practice the *classical three-step method* of estimation is more often used. In its step 1, model (2) without covariates is first estimated. In step 2, this model is used to assign respondents to the latent classes X_{ij} and W_j, conditional on their observed responses $\mathbf{Y}_j$; how this is done for the multilevel LC model is described in detail in Vermunt (2003). In step 3 the assigned latent classes are modelled given covariates, treating the classes now as observed variables. This is straightforward to do. However, it, yields biased estimates of the parameters of the structural model, because the assigned classes are potentially misclassified versions of the true latent classes.

Because of this bias in the classical three-step approach, *bias-adjusted stepwise methods* are needed. One such method for multilevel LC models with covariates is the two-stage estimator proposed by Di Mari et al. (2022) - see also Bakk et al. (2022). It involves the following two stages:

(A) First stage: Unconditional multilevel LC model building (measurement model construction).

Step 1: A single-level latent class model is fitted for $\mathbf{Y}_{ij}$ given the low-level latent class X_{ij}, ignoring the multilevel structure of the data. This gives an initial estimate of $\boldsymbol{\Phi}$.

Step 2.a: The multilevel model without covariates (equation 2) is estimated, keeping $\boldsymbol{\Phi}$ fixed at its estimated value from Step 1. This gives estimates of $\boldsymbol{\omega}$ and $\boldsymbol{\Pi}$.

Step 2.b: The two-level model is estimated again, now keeping $\boldsymbol{\omega}$ and $\boldsymbol{\Pi}$ fixed at their estimates from Step 2.a. This gives the estimate of $\boldsymbol{\Phi}$ which is taken forward to the second stage.

(B) Second stage: Inclusion of covariates in the model (structural model construction).

Step 3: The multilevel model (5) with covariates is estimated, keeping the measurement parameters $\boldsymbol{\Phi}$ fixed at their estimates from the first stage. This gives the two-stage estimates of the structural parameters $\boldsymbol{\theta}_2$.

While effective, the two-stage approach has some shortcomings. Although Steps 2.a and 2.b both estimate only part of the measurement model parameters, computationally they do not save much effort because the most challenging part of the estimation (the E-step of the EM algorithm; see below) is required by both steps. Fixing the response probabilities is also not enough to prevent label switching of the classes from one step to the next in the first stage, since this can simultaneously occur at both the low and high levels. Finally, estimating the correct form of the second-stage information matrix, which should take variability of the previous steps into account, is difficult due to the sequential re-updating of the measurement model. These complications make it desirable to look for more straightforward bias-adjusted stepwise approaches for the multilevel LC model. Such a method, the two-step estimator, is described next.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

3. Two-step Estimator for the Model with Covariates

We propose to amend the two-stage estimator by concentrating all of the measurement modeling into a single step 1, where we estimate the multilevel LC model but without covariates. The estimated parameters of the measurement model for the items $\mathbf{Y}_{ij}$ from this step are then taken forward as fixed to step 2, where the structural model for the latent classes given covariates is estimated. Step 2 is thus the same as the second stage of two-stage estimation, but the three steps of its first stage are here collapsed into the single step 1.

The two-step estimation procedure for multilevel LC models that is described in this section has been implemented in the R package `multilevLCA` (Lyrvall et al., 2023) , which can be downloaded from CRAN. The package's routines have been used for the simulations and data analysis in Sects. 4, and 5 of the paper.

3.1. Step 1 — Measurement Model

In the first step, a simple multilevel LC model without covariates is fitted to the data. Given the data defined above, the log likelihood for this step is

$$\ell_1 = \ell(\boldsymbol{\Phi}, \Pi, \boldsymbol{\omega}) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j), \tag{6}$$

where $P(\mathbf{Y}_j)$ is given by (2). This is maximized to find the ML estimate of the parameters of this model. Direct (numerical) maximization is possible, either with suitable constraints or by adopting well-known logistic re-parametrizations, but it quickly becomes infeasible even for a moderate number of low- and/or high-level classes. A more practical alternative to maximize (6) is by means of the expectation-maximization (EM) algorithm (Dempster et al., 1977) , which is what we propose here.

A standard implementation of EM would involve computing $M \times T^{n_j}$ joint posterior probabilities, which is infeasible already with a few low-level units per high-level unit. Instead, our implementation of the EM algorithm follows closely Vermunt (2003)'s *upward–downward* method of computing the joint posteriors of the low- and high-level classes (see also (Vermunt, 2008)), where the number of joint posterior probabilities to be computed is only a linear function of the number of low-level units per high-level unit. Here we describe in detail the E and M steps of the algorithm, with the step-by-step implementation, that we use to obtain the estimates in Step 1.

Using standard EM terminology, let us introduce the following augmenting variables:

$$u_{j,m} = \begin{cases} 1, & \text{if } W_j = m \\ 0, & \text{otherwise.} \end{cases}, \quad v_{i,j,t,m} = \begin{cases} 1, & \text{if } X_{ij} = t, \quad W_j = m, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Defining the *complete-data* sample as $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_J, v_{1,1}, \ldots, u_{j,m}, \ldots, u_{J,M}, v_{1,1,1,1}, \ldots, v_{i,j,t,m}, \ldots, v_{n_J,J,T,M}\}$, the *complete–data log–likelihood* (CDLL) for the first step can be specified as

$$\begin{aligned} \ell_1^c = & \sum_{j=1}^{J}\sum_{m=1}^{M} u_{j,m} \log(\omega_m) + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} v_{i,j,t,m} \log(\pi_{t|m}) \\ & + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} v_{i,j,t,m} \sum_{h=1}^{H} \{Y_{ijh} \log(\phi_{h|t}) + [1 - Y_{ijh}] \log(1 - \phi_{h|t})\}, \end{aligned} \tag{8}$$

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

where we have dropped the argument $(\boldsymbol{\Phi}, \Pi, \boldsymbol{\omega})$ from ℓ_1^c for simplicity of notation.

In the E step, the missing data are imputed by conditional expectations given the observed data and current values for the unknown model parameters. More specifically, this involves the computation of the following expected CDLL

$$\begin{aligned}\mathbb{E}\left[\ell_1^c\right] = &\sum_{j=1}^{J}\sum_{m=1}^{M}\hat{u}_{j,m}\log(\omega_m) + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T}\hat{v}_{i,j,t,m}\log(\pi_{t|m}) \\ &+ \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T}\hat{v}_{i,j,t,m}\sum_{h=1}^{H}\{Y_{ijh}\log(\phi_{h|t}) + [1-Y_{ijh}]\log(1-\phi_{h|t})\} \equiv Q, \end{aligned} \tag{9}$$

where

$$\hat{u}_{j,m} = \frac{\omega_m \prod_{i=1}^{n_j}\sum_{t=1}^{T}\pi_{t|m}\prod_{h=1}^{H}P(Y_{ijh}|X_{ij}=t)}{\sum_{l=1}^{M}\omega_l\prod_{i=1}^{n_j}\sum_{t=1}^{T}\pi_{t|l}\prod_{h=1}^{H}P(Y_{ijh}|X_{ij}=t)}. \tag{10}$$

To compute the conditional expectation of $v_{i,j,t,m}$, we use the fact that the joint probability $P(X_{ij}=t, W_j=m|\mathbf{Y}_j)$ can be written as $P(W_j=m|\mathbf{Y}_j)P(X_{ij}=t|W_j,\mathbf{Y}_j)$, where $P(W_j=m|\mathbf{Y}_j)$ is already available from (10). Note that, given the model assumptions,

$$P(X_{ij}=t|W_j,\mathbf{Y}_j) = P(X_{ij}=t|W_j,\mathbf{Y}_{ij}), \tag{11}$$

which we use to compute the following desired quantity

$$\begin{aligned}\hat{v}_{i,j,t,m} &= P(X_{ij}=t, W_j=m|\mathbf{Y}_j) \\ &= P(W_j=m|\mathbf{Y}_j)P(X_{ij}=t|W_j,\mathbf{Y}_{ij}) \\ &= \hat{u}_{j,m}\frac{P(X_{ij}=t|W_j=m)P(\mathbf{Y}_{ij}|X_{ij}=t)}{P(\mathbf{Y}_{ij})} \\ &= \hat{u}_{j,m}\frac{\pi_{t|m}\prod_{h=1}^{H}P(Y_{ijh}|X_{ij}=t)}{\sum_{s=1}^{T}\pi_{s|m}\prod_{h=1}^{H}P(Y_{ijh}|X_{ij}=s)}, \end{aligned} \tag{12}$$

where in the third row we are using the assumption that the joint probability function of the response variables depend on high–level class membership only through low-level class membership. For the unrestricted multi–group LC model, the expression (12) would be adapted straightforwardly.

In the M step of the algorithm, the expected CDLL (9) is maximized with respect to the model parameters $(\boldsymbol{\Phi}, \Pi, \boldsymbol{\omega})$ subject to the usual *sum–to–one* constraints on probabilities. This yields the following closed–form updates

$$\omega_m = \frac{\sum_{j=1}^{J}\hat{u}_{j,m}}{\sum_{j=1}^{J}\sum_{m=1}^{M}\hat{u}_{j,m}}, \tag{13}$$

$$\pi_{t|m} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\hat{v}_{i,j,t,m}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{t=1}^{T}\hat{v}_{i,j,t,m}}, \tag{14}$$

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

$$\phi_{h|t} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\hat{v}_{i,j,t,m}Y_{ijh}}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\hat{v}_{i,j,t,m}}. \tag{15}$$

Starting from initial values for the model parameters, the algorithm iterates between the E- and the M-steps until some convergence criterion is met, e.g. until the difference between the log-likelihood values of two subsequent iterations falls below some threshold value.

As for all mixture models, the log-likelihood function can have several local optima and there is no guarantee that the solution found by the EM algorithm is the global optimum (Wu, 1983) . To better explore the likelihood surface, multiple starting value strategies are typically implemented (among others, see (Biernacki et al., 2003; Maruotti & Punzo, 2021)). Beyond doubt, the easiest, and most common approach is to initialize the EM algorithm randomly from several different starting points. However, even for relatively simpler models, the multiple starting value strategy is often outperformed by more refined techniques (Biernacki et al., 2003) .

For any stepwise estimators, the initialization strategy of earlier steps is particularly relevant because subsequent steps will be conditional on estimates from previous steps. In our step 1, we suggest implementing the following hierarchical initialization strategy (for a similar approach in a related context, see for instance (Catania & Di Mari, 2021; Catania et al., 2022)):

(1) Perform a single–level K–modes clustering (Huang, 1997; MacQueen, 1967) , with $K = M$. For each $j = 1, \ldots, J$
   - let $\dot{W}_{ij}$ be the outcome class assignment for unit i in group j;
   - specify $\widetilde{W}_j$ as the most frequent assigned class among the n_j observations belonging to group j, and let $\widetilde{W}_{ij} = \widetilde{W}_j$ for all $i = 1, \ldots, n_j$.

   The relative sizes of the resulting high–level classes are used to initialize $\boldsymbol{\omega}$. The entries of $\boldsymbol{\omega}$, before being carried over to the actual estimation step, can be sorted in increasing or decreasing order.

(2) Fit a single–level T–class LC model on the pooled data, ignoring the multilevel structure. Note that the K–modes algorithm can be employed herein as well to initialize the single–level LCA. The estimated output is organized as follows
   - the response probabilities are passed on the EM algorithm as a start for $\boldsymbol{\Phi}$;
   - let $\widetilde{X}_{ij}$ be the maximum a posteriori class assignment for unit i in group j. Cross–tabulate $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{W}}$, where $\widetilde{\mathbf{X}} = (\widetilde{X}_{11}, \ldots, \widetilde{X}_{n_J J})'$, and $\widetilde{\mathbf{W}} = (\widetilde{W}_{11}, \ldots, \widetilde{W}_{n_J J})'$. From the $T \times M$ table of joint counts, compute the conditional (relative) counts of $\widetilde{\mathbf{X}}|\widetilde{\mathbf{W}}$ to initialize Π.

   The low-level classes can be re-ordered by letting low-level cluster 1 be the one with the highest average probability to score a "1" on all items, cluster 2 the one with the second highest average probability to score a "1" on all items, and so on.

Note that the suggested rule to re-order low-level classes is only an example of a rule that is often (but not always) useful. This is because, if there are many items or some are for rare characteristics, the joint probability of scoring "1" on all of them together might be a number so small as to be overwhelmed by sampling error or even by machine imprecision. That would effectively bring label switching back again. In cases like these, we suggest implementing alternative re-ordering principles.

Running the EM algorithm to convergence from the above starting values, the solution with the highest log-likelihood (6) provides us with estimates $\hat{\boldsymbol{\omega}}$, $\hat{\Pi}$, $\hat{\boldsymbol{\Phi}}$. Of these, $\hat{\boldsymbol{\omega}}$ and $\hat{\Pi}$ are discarded and $\text{vec}(\hat{\boldsymbol{\Phi}}) = \hat{\boldsymbol{\theta}}_1$ are retained as the estimates of the measurement parameters $\boldsymbol{\theta}_1$ from this step 1.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

3.2. Step 2 — Model for Class Membership

In the second step of estimation, the parameters $\boldsymbol{\theta}_2$ of the model for the latent classes in Eq. (5) are estimated, keeping the measurement parameters $\boldsymbol{\theta}_1$ fixed at their step-1 estimates $\hat{\boldsymbol{\theta}}_1$ (see Fig. 2). These step-2 estimates are obtained by maximizing the pseudo log-likelihood function

$$\ell_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j|\mathbf{Z}_j) \tag{16}$$

with respect to $\boldsymbol{\theta}_2$. Here $\log P(\mathbf{Y}_j|\mathbf{Z}_j)$ is given by Eq. (5), except that $\hat{\boldsymbol{\theta}}_1$ are regarded as fixed and known values rather than unknown parameters. The EM algorithm that we propose for this step works similarly to the one that we used for the first step. In particular, under the definition of the augmenting variables given in Sect. 3.1, the CDLL is given by

$$\begin{aligned}\ell_2^c = &\sum_{j=1}^{J}\sum_{m=1}^{M} u_{j,m} \log(\omega_m) + \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} v_{i,j,t,m} \log\left(\frac{\exp(\boldsymbol{\gamma}'_{tm}\mathbf{Z}_{ij})}{1+\sum_{s=2}^{T}\exp(\boldsymbol{\gamma}'_{tm}\mathbf{Z}_{ij})}\right) \\ &+ \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} v_{i,j,t,m} \sum_{h=1}^{H}\{Y_{ijh}\log(\hat{\phi}_{h|t}) + [1-Y_{ijh}]\log(1-\hat{\phi}_{h|t})\},\end{aligned} \tag{17}$$

where we have dropped the argument $(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1)$ from ℓ_2^c for ease of notation. Note that the E step is analogous as that described in Sect. 3.1, except that now the low-level class probabilities conditional on high-level membership depend on covariates. In the M step the expected CDLL, obtained by substituting the missing values with expectations computed using analogous formulas as (10) and (12), is maximized with respect to $\boldsymbol{\theta}_2$ only. Whereas the update for $\boldsymbol{\omega}$ is given by (13), to derive the update for the regression coefficients note that $v_{i,j,t,m} = P(X_{ij} = t, W_j = m|\mathbf{Y}_j)$ can be written as the product of $u_{j,m} = P(W_j = m|\mathbf{Y}_j)$ and $q_{i,j,t|m} = P(X_{ij} = t|W_j, \mathbf{Y}_j)$. Thus, estimates of $\boldsymbol{\Gamma}$ can be found solving the equations

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} \hat{u}_{j,m}\hat{q}_{i,j,t|m} \frac{\partial \log\left(P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij})\right)}{\partial \text{vec}(\boldsymbol{\Gamma})} = 0, \tag{18}$$

which are weighted sums of M equations, each with weights $\hat{q}_{i,j,t|m}$.

Stepwise estimation is well known to enhance algorithm stability and speed of convergence (Bakk & Kuha, 2018; Bartolucci et al., 2015; Di Mari & Maruotti, 2022; Skrondal & Kuha, 2012) . However, class labels in multiple hidden layer models can still be switched, and keeping the response probabilities fixed cannot prevent it as there are still $M!$ possible permutations of the high-level class labels. We handle this issue by initializing $\boldsymbol{\omega}$ at its estimate from the first step, and by taking $\log\left(\pi_{t|m}/\pi_{1|m}\right)$ to initialize the intercepts γ_{0tm}, for all $m = 1, \ldots, M$ and $t = 2, \ldots, T$. The other elements of $\boldsymbol{\Gamma}$ are initialized at zero.

3.3. Selecting the Number Latent Classes

The description of the two-step estimation procedure above takes the numbers of latent classes at both the lower and higher levels as given. The selection of these numbers is a separate exercise. It is normally carried out without covariates, and the selected numbers of classes are then held

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

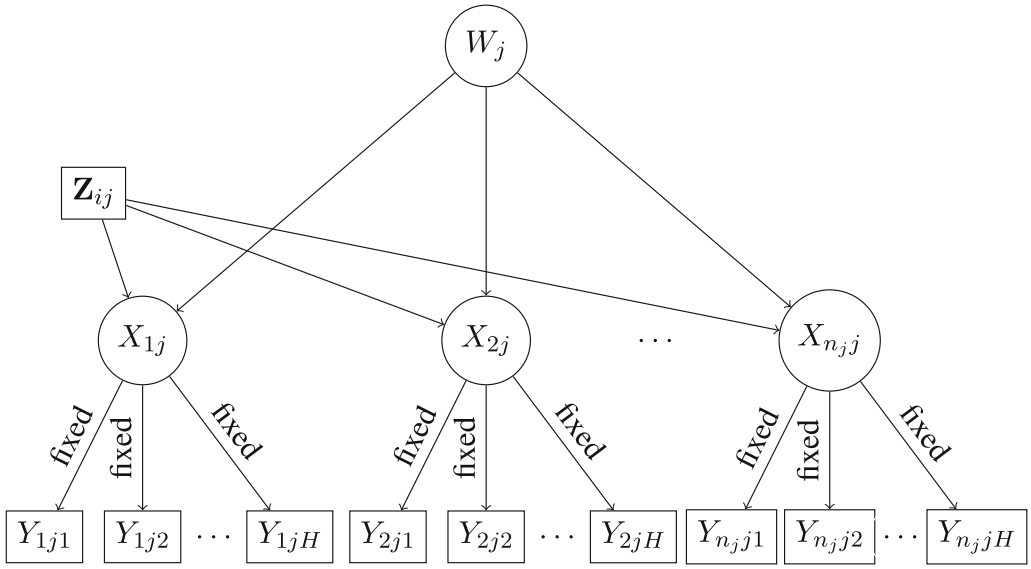


FIGURE 2.
Step 2 of the two-step estimation: Estimating the structural model for low-level latent classes X_{ij} given covariates $\mathbf{Z}_{ij}$ and high-level latent classes W_j, keeping measurement model parameters for items Y_{ijh} fixed at their estimates from Step 1.

fixed when covariates are added. This is also in line with general recommendations for LCA with covariates (Masyn, 2017) .

The selection of the numbers of classes could be considered as a joint exercise of both the high and low levels together, but a generally used recommendation is to use instead a hierarchical procedure which selects them one at a time (Lukociene et al., 2010) . First, simple LC models are fitted at the lower level and the number of classes for it (T) is selected. Second, this number is held fixed, and multilevel LC models are fitted and compared to select the number of classes at the higher level (M). Third, the selected M is fixed, and model selection for the multilevel model is done again at the lower level, to obtain the final value of T. A still simpler approach would skip the third step (Vermunt, 2003) , but including it allows us to check if the selected number of lower-level classes changes once the within-group associations induced by the high-level classes are allowed for.

This hierarchical approach can be used with any method of estimating the models. However, when combined with our two-step estimator, simultaneously selecting the number of classes of the measurement at both levels is also feasible. Practically, this is possible by leveraging an efficient integration of the above initialization strategy with parallel (multi-core) estimation of all plausible values of T and M.

The best candidate values of M and T can be selected with standard information criteria, like AIC or BIC. For the final choice, we suggest balancing the use information criteria with the evaluation of low- and high-level class separation, and, perhaps most importantly, the substantive inspection of the candidate model configurations. For a wider discussion on this issue, see, among others, Di Mari et al. (2022); Magidson and Vermunt (2004). In the social sciences, one of the most commonly used measures of class separation is the entropy-based R^2 of Magidson (1981). The latter can be defined at both lower and higher levels to judge class separation (see (Di Mari et al., 2022; Lukociene et al., 2010)).

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

3.4. Statistical properties of the two-step estimator

Our two-step estimator is an instance of pseudo maximum likelihood estimation (Gong and Samaniego, 1981) . Such estimators are consistent and asymptotically normally distributed under very general regularity conditions. The conditions and a proof of consistency can be found in (Gourieroux and Monfort (1995), Sec. 24.2.4). Let the true parameter vector be $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_1^{\star\prime}, \boldsymbol{\theta}_2^{\star\prime})'$. If the one-step ML estimator of $\boldsymbol{\theta}$ is itself consistent for $\boldsymbol{\theta}^\star$, in order to prove consistency of our two-step estimator $\hat{\boldsymbol{\theta}}$ it suffices to show that (1) $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can vary independently of each other, and (2) $\hat{\boldsymbol{\theta}}_1$ is consistent for $\boldsymbol{\theta}_1^\star$. These conditions are satisfied in our case: (1) is true by construction of the model, and (2) is satisfied since $\hat{\boldsymbol{\theta}}_1$ from step 1 is a ML estimate of the measurement model parameters of the multilevel LC model without covariates, and these parameters are taken to be the same as in the model with covariates.

Let $\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denote the joint log-likelihood function for the model, let $\bar{\mathbf{s}}_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)$ denote the mean score $N^{-1}\partial\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2$ evaluated at $(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)$, where N denote the overall sample size, and let

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*) = \begin{bmatrix} \boldsymbol{\mathcal{I}}_{11} \\ \boldsymbol{\mathcal{I}}_{21} \; \boldsymbol{\mathcal{I}}_{22} \end{bmatrix},$$

be the Fisher information matrix. In addition, let us suppose that

$$N^{1/2} \begin{bmatrix} \hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star \\ \bar{\mathbf{s}}_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star) \end{bmatrix} \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} \\ \boldsymbol{\Sigma}_{21} \; \boldsymbol{\mathcal{I}}_{22} \end{bmatrix}\right).$$

Then, using the results of Theorem 2.2 of Gong and Samaniego (1981) (see also (Parke, 1986)),

$$N^{1/2}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^\star) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{V}), \tag{19}$$

where $\hat{\boldsymbol{\theta}}_2$ is the proposed two-step estimator and

$$\boldsymbol{V} = \underbrace{\boldsymbol{\mathcal{I}}_{22}^{-1}}_{\equiv \boldsymbol{V}_2} + \underbrace{\boldsymbol{\mathcal{I}}_{22}^{-1}\boldsymbol{\mathcal{I}}_{21}\boldsymbol{\Sigma}_{11}\boldsymbol{\mathcal{I}}_{21}'\boldsymbol{\mathcal{I}}_{22}^{-1}}_{\equiv \boldsymbol{V}_1}. \tag{20}$$

Intuitively, $\boldsymbol{V}_2$ describes the variability in $\hat{\boldsymbol{\theta}}_2$ given the step one estimates $\hat{\boldsymbol{\theta}}_1$, and $\boldsymbol{V}_1$ the additional variability arising from the fact that $\boldsymbol{\theta}_1$ are not known but rather estimated by $\hat{\boldsymbol{\theta}}_1$ with their own sampling variability.

Let $\mathbf{s}_{ij,\boldsymbol{\theta}_2}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ be the individual contribution to the score of low-level unit i belonging to high-level group j evaluated at the parameter estimates of the first and second step respectively. To compute such score we use the well-known fact that $\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \partial Q/\partial\boldsymbol{\theta}$ (Oakes, 1999) , where $Q = \mathbb{E}\left[\ell^c(\boldsymbol{\theta})\right]$. All such quantities are available from the above EM algorithm without any extra effort. Therefore, $\boldsymbol{\mathcal{I}}_{22}$ and $\boldsymbol{\mathcal{I}}_{21}$ can be estimated respectively as

$$\hat{\boldsymbol{\mathcal{I}}}_{22} = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{s}_{ij,\boldsymbol{\theta}_2}(\hat{\boldsymbol{\theta}}_2)\,\mathbf{s}_{ij,\boldsymbol{\theta}_2}(\hat{\boldsymbol{\theta}}_2)' \tag{21}$$

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

and

$$\hat{\mathcal{I}}_{21} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \mathbf{s}_{ij,\boldsymbol{\theta}_2}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)\, \mathbf{s}_{ij,\boldsymbol{\theta}_1}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)'. \tag{22}$$

An estimate $\hat{\Sigma}_{11}$ can be obtained analogously by fitting model (2). We give details on the derivations of the desired quantities in the appendix.

Note that Equation (20) shows that there is a loss of efficiency of the two-step estimator with respect to the simultaneous ML estimator. This important theoretical and practical aspect with be investigated in the simulation study—although we expect this loss to be rather small as very little information about $\boldsymbol{\theta}_2$ should be contained in $\mathbf{Y}$.

4. Simulation Study

4.1. Settings

We conduct a simulation study to investigate the finite sample properties of the proposed two-step estimator. It is compared with the simultaneous (one-step) estimator and the two-stage estimator of Bakk et al. (2022); Di Mari et al. (2022). One-step estimation is the statistical benchmark, and the two-step estimator's performance is evaluated in terms of its statistical and computational performance relative to this benchmark. The target measures that we use for the comparison are the bias, standard deviations, confidence interval coverage rates, and computation time of the stepwise estimators compared with those of the simultaneous estimator. We compute both absolute standard deviations, to assess the efficiency of our estimator, as well as relative standard deviations with respect to the one-step method, to investigate potential loss of efficiency with respect to the benchmark. Class separation and sample size are well-known determinants of the finite-sample behavior of stepwise estimators for LCA (Bakk & Kuha, 2018; Vermunt, 2010) . We considered all combinations of larger and smaller sample sizes, at higher level (30, 50, or 100 higher-level units) and lower level (100 or 500), with a total of 6 sample size conditions. Data were generated from a multilevel LC model with 2 high-level classes and 3 low-level classes and with 10 binary indicators and one continuous covariate generated from a standard normal distribution. The random slopes $\gamma_{2|1}$, and $\gamma_{3|1}$ were set to -0.25 and -0.25, whereas $\gamma_{2|2}$, and $\gamma_{3|2}$ to 0.25 and 0.25, corresponding to a moderate magnitude on the logistic scale.

In multilevel LC models, separation plays a role at both low and high levels (Lukociene et al., 2010) . We manipulate low-level class separation by allowing the the response probabilities for the most likely responses to be either 0.7, 0.8 or 0.9, corresponding respectively to low, moderate, and large class separation. We remark that the low class separation condition can be considered as an extreme scenario, in which LCA is hardly carried out in practice. Nevertheless, we decide to include it as a benchmarking condition. Class profiles are such that the first class has high probability to score 1 on all items, the second class to score 1 on the last five items and 0 on the first 5 items, and the third class is likely to score 0 on all items. At the high level, in the model for W, we manipulate class separation by altering the random intercept magnitudes, which are both relatively close to zero in the moderate separation case (-0.85, -1.38 and 0.85, 1.38), and further away from zero in the large separation case (-1.38, -2.07 and 1.38, 2.07). These simulation conditions are in line with previous studies on multilevel LCA (Lukociene et al., 2010; Park & Yu, 2018) .

We generated 500 samples for each of the 36 crossed simulation factors of low-level and high-level sample size and low-level and high-level class separation (see Table 1). Data generation and

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

TABLE 1.
24 simulation conditions.

Condition	LL sample size	HL sample size	LL separation	HL separation
1	100	30	Small	Moderate
2	500	30	Small	Moderate
3	100	50	Small	Moderate
4	500	50	Small	Moderate
5	100	100	Small	Moderate
6	500	100	Small	Moderate
7	100	30	Moderate	Moderate
8	500	30	Moderate	Moderate
9	100	50	Moderate	Moderate
10	500	50	Moderate	Moderate
11	100	100	Moderate	Moderate
12	500	100	Moderate	Moderate
13	100	30	Large	Moderate
14	500	30	Large	Moderate
15	100	50	Large	Moderate
16	500	50	Large	Moderate
17	100	100	Large	Moderate
18	500	100	Large	Moderate
19	100	30	Small	Large
20	500	30	Small	Large
21	100	50	Small	Large
22	500	50	Small	Large
23	100	100	Small	Large
24	500	100	Small	Large
25	100	30	Moderate	Large
26	500	30	Moderate	Large
27	100	50	Moderate	Large
28	500	50	Moderate	Large
29	100	100	Moderate	Large
30	500	100	Moderate	Large
31	100	30	Large	Large
32	500	30	Large	Large
33	100	50	Large	Large
34	500	50	Large	Large
35	100	100	Large	Large
36	500	100	Large	Large

LL stands for Low-Level, HL stands for High-Level.

model estimation were carried out in R (Venables et al., 2013), with the integration of C++ code for computation efficiency (Eddelbuettel & François, 2011).

4.2. *Results*

All estimators show very similar values for bias (see Figs. 3a, b), and both two-stage and two-step estimators have nearly identical results compared to the simultaneous estimator. Relative efficiency with respect to the simultaneous estimator (Table 8, in the appendix) is, in all conditions, approximately one for both stepwise estimators, with the two-stage estimator doing very slightly worse only in one condition. Confidence interval coverages (Fig. 4) are mostly very

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

similar between the three estimators. We observe some undercoverage for all methods in the low-separation and small high-level sample size conditions. This may be due to the fact that expected information matrices are used to estimate the asymptotic variance covariance matrix, rather than the observed ones, and the contributions to the score are computed on high level units, and to the overlap between classes.

The different estimators thus perform essentially identically. Where they differ from each other is in their computational demands. Considering the computation time relative to the simultaneous estimator (Fig. 5), we find that both stepwise estimators are always (and up to four times) faster than the simultaneous estimator, and the two-step estimator achieves this with one fewer step compared to the existing two-stage competitor.

5. Analysis of Cross-National Citizenship Norms with Multilevel LCA

In this empirical example, we analyze citizenship norms in a diverse set of countries. The data are taken from the International Civic and Citizenship Education Study (ICCS) conducted by the International Association for the Evaluation of Educational Achievement (IEA). Prior research has used LCA to analyze the first two waves of this survey, which were conducted in 1999 and 2009, to investigate distinctive types of citizenship norms (Hooghe & Oser, 2015; Hooghe et al., 2016; Oser & Hooghe, 2013) . We focus on the most recent round of the survey, from 2016 (Köhler et al., 2018) . The data are from a survey of students in their eighth year of schooling. We have data from between 1300 and 7000 respondents in each of 24 countries, as shown in Table 2.

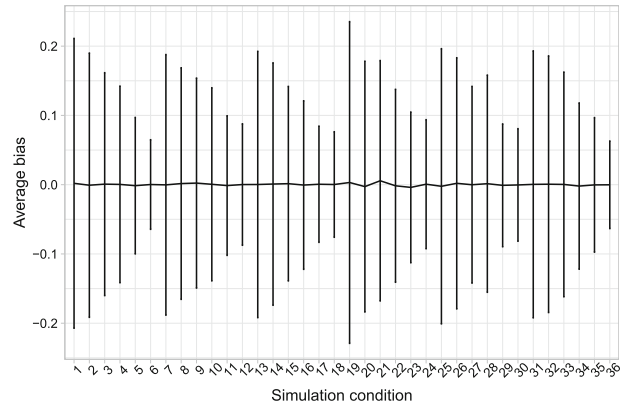
The respondents answered 12 questions (items) on how important they think different behaviours are for ”being a good adult citizen”. These behaviours were always obeying the law (labelled *obey* below), taking part in activities promoting human rights (*rights*), participating in activities to benefit people in the local community (*local*), working hard (*work*), taking part in activities to protect the environment (*envir*), voting in every national election (*vote*), learning about the country's history (*history*), showing respect for government representatives (*respect*), following political issues in the newspaper, on the radio, on TV or on the Internet (*news*), participating in peaceful protests against laws believed to be unjust (*protest*), engaging in political discussions (*discuss*), and joining a political party (*party*).

We treat these twelve items as indicators of the individuals' perceptions of the duties of a citizen (*citizenship norms*). The data have a multilevel structure, with individuals as the low-level units and countries as the high-level units. As predictors of low-level latent class membership, we include the respondent's gender, socio-economic status operationalised by the proxy measure of the number of books in their home, and measures of the respondent's educational expectations, parental education, and if she/he is a non-native language speaker. For details on data cleaning and recoding, see Oser et al. (2023).

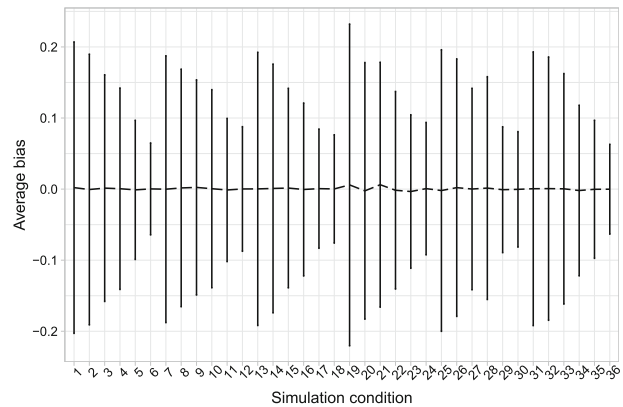
To compare with previous work on the same data, we fit a multilevel LC model with $T = 4$ low-level classes (of individuals within countries) and $M = 3$ high-level classes (of countries). The same data set was analyzed in Di Mari et al. (2022) with a multilevel LC model with random intercepts, estimated with a two-stage estimator. We extend Di Mari et al. (2022)'s model specification by allowing for both random intercepts and random slopes, and we fit the model with the proposed two-step estimator. As the two-step estimator has been shown to be computationally more efficient than the two-stage estimator though with equal performances, for the comparison we include the benchmark simultaneous estimator only.

The measurement model, at both levels, presents very well separated classes (Table 3). At the lower level, the four latent classes are characterised by their the conditional response probability patterns, as shown in Fig. 6. Two classes present response configurations relating to two relevant and well-known notions of citizenship norms. First, a “Duty” group, which places high

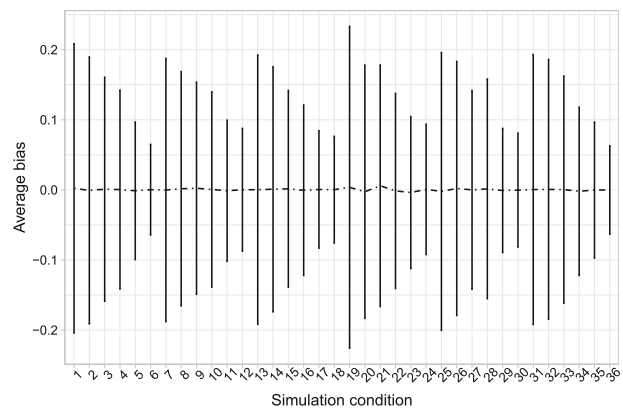
https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press



(a) One–step estimator



(b) Two–stage estimator



(c) Two–step estimator

FIGURE 3.
Line graphs of estimated bias for the one-step, two-step, and two-stage estimators, for the 36 simulation conditions, averaged over the 500 replicates. Error bars are based on mean bias ± Monte Carlo standard deviations.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

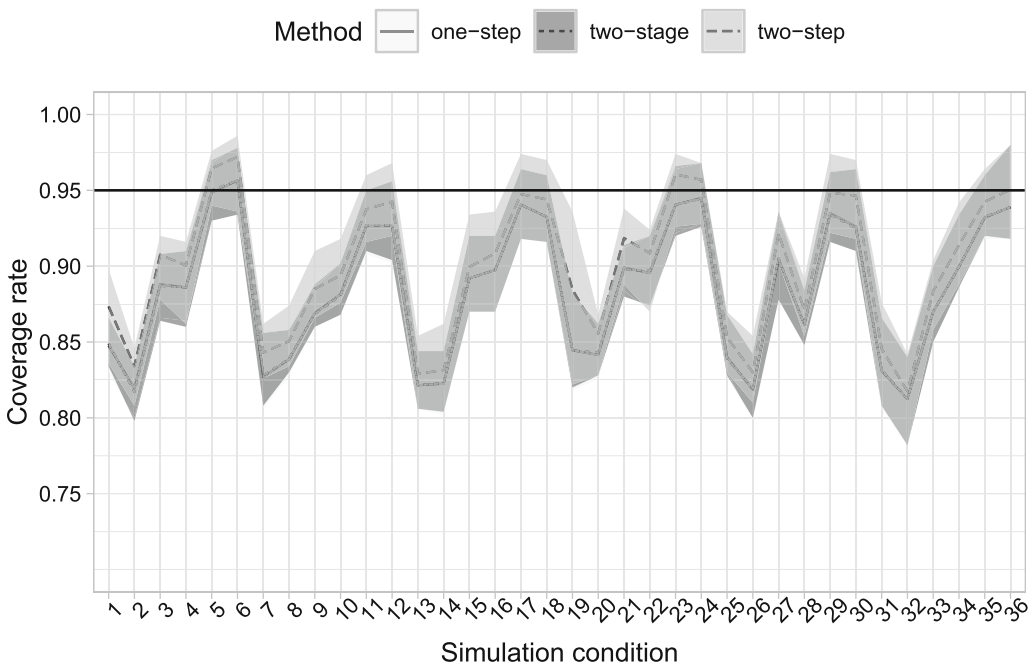


FIGURE 4.
Observed coverage rates of 95% confidence intervals, averaged over covariate effects, for the one-step, two-stage and two-step estimators for the 36 simulation condition, averaged over the 500 replicates. Lower and higher confidence values reported in the confidence bars, based on the minimum and maximum coverages of the confidence intervals for each covariate effect.

importance on the act of voting, discussing politics, and party activity, while manifesting relatively low interest in protecting human rights and activities to assist the local community. Second, an "Engaged" group, which displays higher emphasis on engaged attitudes such as protecting the environment, and lower importance on more traditional citizenship activity items such as membership of political parties. In addition, we observe two classes with consistently high and consistently low probabilities of assigning importance to all of the behaviours, here labelled the "Maximal" and the "Subject" classes respectively.

At the higher level, the estimated model includes three latent classes for the countries, labelled below as HL1, HL2 and HL3. Considering first the conditional probabilities for the four individual-level classes given these country-level classes (see Table 4), we can see that HL1 has clearly the highest conditional probability for the individual "Duty" class, HL2 for the "Maximal" class and HL3 for the "Engaged". The country classes do not differ in probabilities of the passive "Subject" class of individuals, which are in any case consistently low. Table 5 shows the assignment of countries to the classes, when the assignment is done based on highest posterior probabilities given the survey responses in the countries. Here there are no very clear patterns. Only two countries (Denmark and Netherlands) are assigned to HL1, while the other two classes each include a fairly heterogeneous subset of the rest of the countries.

Table 6 presents estimates of the parameters of main interest in the analysis, the coefficients of the structural model for the lower-level classes given individual-level covariates, separately within each of the higher-level classes. We note first that the one-step and two-step estimates and their standard errors are very similar, as would be expected given the previous simulation results.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

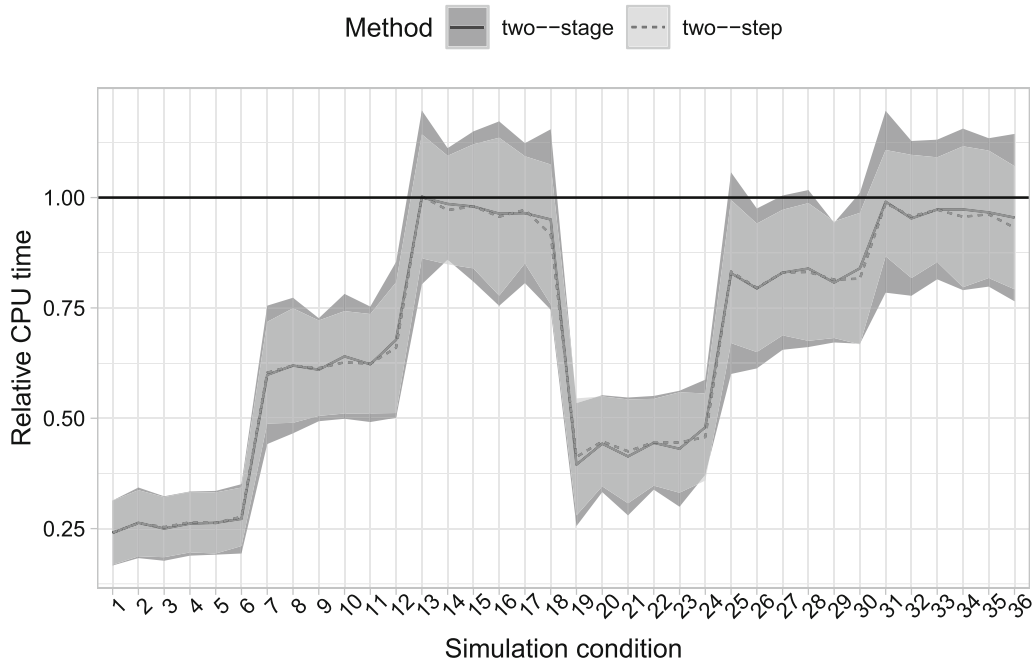


FIGURE 5.
Relative computation time for the one-step, two-stage and two-step estimators for the 24 simulation condition, averaged over the 500 replicates. The one-step estimator's estimation time is taken as reference. Confidence bands based on average values ± their Monte Carlo standard deviation.

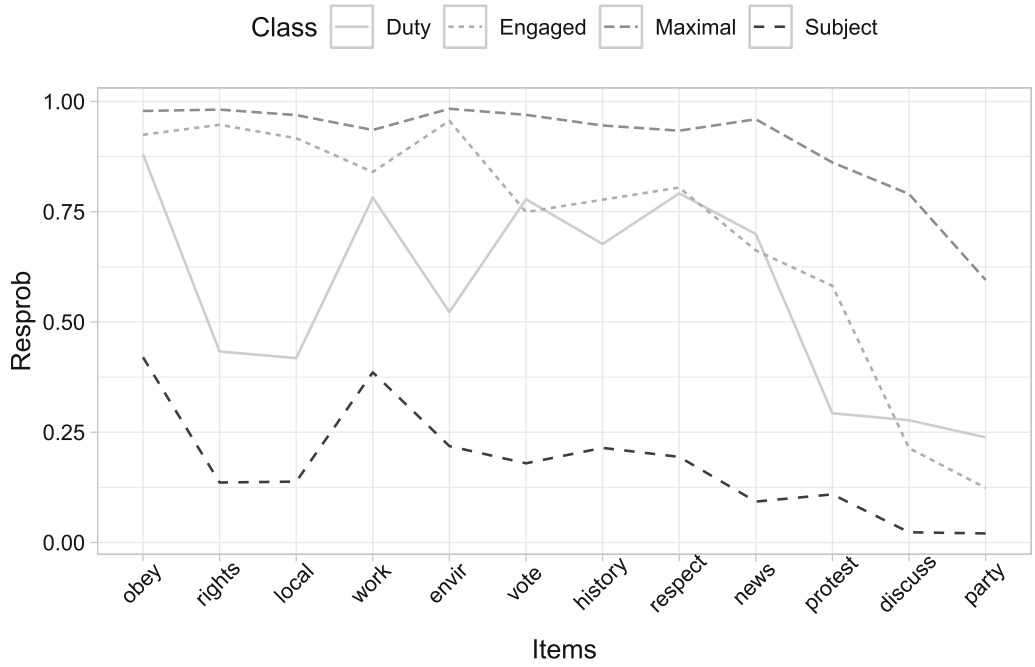


FIGURE 6.
Measurement model at the lower (individual) level: line graph of the class-conditional response probabilities.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

TABLE 2.
Number of respondents per country of the third wave (2016) of the IEA survey used for the analysis.

Country	Sample size
Belgium	2750
Bulgaria	2682
Chile	4753
Colombia	4992
Denmark	5692
Germany	1313
Dominican Republic	2779
Estonia	2770
Finland	3037
Hong Kong	2553
Croatia	3655
Italy	3274
Republic of Korea	2557
Lithuania	3422
Latvia	3000
Mexico	4987
Malta	3317
Netherlands	2692
Norway	5740
Peru	4713
Russia	7049
Slovenia	2664
Sweden	2828
Taiwan	3904

TABLE 3.
Summary statistics for the measurement model.

	Value
log-likelihood	−459295.5
BIC	919262.1
BIC (J)	918778.5
$\text{entrR}^2_{\text{low}}$	0.999
$\text{entrR}^2_{\text{high}}$	0.999
npar	59

Considering the coefficients themselves, note that they compare each of the other classes to the "Maximal" class for whom all of the behaviours are to a greater or less extent considered important to good citizenship. Compared to this class, the relative probability of the (overall quite small) "Subject" class for whom none of the behaviours are important, is higher for individuals who are boys, speak the native language at home, have fewer books at home, and have low educational aspirations. The probabilities of the "Engaged" class, who are partly similar to "Maximal" but place less importance on many of the traditional political activities, are relatively higher for girls, those who have larger number of books at home, and for native speakers. For the "Duty"

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

TABLE 4.
Estimated proportions of low-level (individual-level) classes conditional on high-level (country-level) class membership.

	HL 1	HL 2	HL 3
Maximal	0.207	0.576	0.317
Engaged	0.290	0.277	0.478
Subject	0.031	0.029	0.044
Duty	0.471	0.118	0.161

TABLE 5.
Assignment of countries to the high-level classes, based on the maximum a posteriori (MAP) classification rule. $M = 3$.

Country	HL 1	HL 2	HL 3
Belgium	0	0	1
Bulgaria	0	0	1
Chile	0	0	1
Colombia	0	0	1
Denmark	1	0	0
Germany	0	0	1
Dominican Republic	0	1	0
Estonia	0	0	1
Finland	0	0	1
Hong Kong	0	1	0
Croatia	0	1	0
Italy	0	1	0
Republic of Korea	0	1	0
Lithuania	0	0	1
Latvia	0	0	1
Mexico	0	1	0
Malta	0	0	1
Netherlands	1	0	0
Norway	0	0	1
Peru	0	1	0
Russia	0	1	0
Slovenia	0	0	1
Sweden	0	0	1
Taiwan	0	1	0

class, which differs from the "Engaged" in placing much less importance on direct activism, the probabilities relative to "Maximal" are higher for boys and those with low educational aspirations. For the comparisons of other pairs of classes, these estimates also imply, for example, that the probabilities of "Engaged" relative to "Duty" are generally higher for girls than for boys. These patterns of the coefficients are broadly similar in each of the country classes, with some variation in detail.

Finally, we report CPU time of estimation and the number of iterations until convergence for the two approaches (Table 7). In this real-data example, the two-step estimator takes only about 22 s to reach convergence, with 26 EM iterations. The one-step estimator requires 261 iterations and a running time of around 4.5 min to reach convergence. Each iteration requires about 0.93 s

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

TABLE 6.
Estimated coefficients of structural models, i.e. multinomial logistic models for membership of the four individual-level latent classes conditional on covariates, separately within each of the three country-level latent classes (HL1, HL2 and HL3).

	Engaged		Subject		Duty	
	One-step	Two-step	One-step	Two-step	One-step	Two-step
HL 1						
Intercept	0.875***	0.944***	0.923***	0.757***	0.945***	0.934***
	(0.009)	(0.009)	(0.010)	(0.010)	(0.159)	(0.156)
Female	0.359***	0.338***	−0.983***	−1.072***	0.140	0.106
	(0.092)	(0.090)	(0.053)	(0.052)	(0.082)	(0.080)
Number of books	−0.016	−0.014	−0.36***	−0.345***	−0.166	−0.173
	(0.080)	(0.079)	(0.080)	(0.079)	(0.175)	(0.171)
Education goal	0.018	0.013	−0.819***	−0.865***	0.232**	0.207
	(0.212)	(0.228)	(0.181)	(0.202)	(0.088)	(0.095)
Mother education	−0.308**	−0.311	−0.314	−0.327	−0.007	−0.002
	(0.116)	(0.124)	(0.135)	(0.148)	(0.133)	(0.143)
Father education	−0.108	−0.117	−0.143	−0.131	−0.164	−0.164
	(0.256)	(0.294)	(0.134)	(0.134)	(0.073)	(0.072)
Non-native language level	−0.437***	−0.428***	−0.03	−0.155	−0.446***	−0.408***
	(0.042)	(0.042)	(0.068)	(0.067)	(0.065)	(0.065)
HL 2						
Intercept	−0.760***	−0.749***	−1.404***	−1.503***	−1.076***	−1.099***
	(0.064)	(0.064)	(0.140)	(0.139)	(0.072)	(0.073)
Female	0.199***	0.180***	−0.651***	−0.672***	−0.255***	−0.278***
	(0.036)	(0.036)	(0.023)	(0.023)	(0.036)	(0.036)
Number of books	−0.133***	−0.130***	−0.247***	−0.265***	−0.090	−0.087
	(0.029)	(0.029)	(0.029)	(0.029)	(0.072)	(0.071)
Education goal	0.025	0.014	−0.536***	−0.555***	−0.306***	−0.313***
	(0.105)	(0.111)	(0.079)	(0.084)	(0.042)	(0.045)
Mother education	0.030	0.035	0.090	0.088	0.191**	0.188**
	(0.056)	(0.059)	(0.060)	(0.064)	(0.060)	(0.064)
Father education	0.018	0.016	−0.160	−0.166	0.022	0.018
	(0.157)	(0.166)	(0.078)	(0.079)	(0.045)	(0.045)
Non-native language level	−0.127***	−0.114***	−0.306***	−0.338***	0.299***	0.290***
	(0.027)	(0.027)	(0.040)	(0.040)	(0.037)	(0.037)
HL 3						
Intercept	0.218***	0.260***	−0.044	−0.217**	−0.040	−0.019
	(0.037)	(0.037)	(0.076)	(0.077)	(0.071)	(0.072)
Female	0.301***	0.282***	−0.587***	−0.616***	−0.230***	−0.261***
	(0.032)	(0.032)	(0.019)	(0.019)	(0.035)	(0.034)
Number of books	−0.083**	−0.081**	−0.358***	−0.374***	−0.083	−0.094
	(0.027)	(0.027)	(0.027)	(0.026)	(0.059)	(0.058)
Education goal	0.148	0.124	−0.547***	−0.544***	−0.411***	−0.434***
	(0.099)	(0.106)	(0.063)	(0.067)	(0.035)	(0.037)
Mother education	0.040	0.044	−0.033	−0.033	0.183***	0.176***
	(0.050)	(0.053)	(0.048)	(0.051)	(0.048)	(0.051)
Father education	−0.097	−0.097	−0.125	−0.125	0.037	0.038
	(0.099)	(0.106)	(0.078)	(0.079)	(0.040)	(0.041)
Non-native language level	−0.426***	−0.414***	−0.107**	−0.095	−0.006	0.004
	(0.023)	(0.023)	(0.039)	(0.039)	(0.036)	(0.036)

The "Maximal" class is taken as the reference level for the response class. The number of books available in the respondent's home is treated as a proxy for the respondent's socio–economic status. Both simultaneous (one-step) and the proposed two-step estimators of the same parameters are shown, with standard errors in parentheses.
***p-value<0.01, **p-value<0.05, *p-value<0.1.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

TABLE 7.
CPU time to estimation in seconds, and number of iterations until convergence for the two methods - one-step and two-step estimators.

	CPU time (in seconds)	Number of iterations until convergence
One-step	242.89	261
Two-step	22.01	26

to run for the one-step estimator, while the two-step estimator uses 0.85 s and much fewer EM iterations overall.

6. Discussion

In this paper we proposed a two-step estimator for the multilevel latent class model with covariates. It concentrates the estimation of the measurement model in a single first step. In the second step, covariates are added to the model, keeping the measurement model parameters fixed. The approach represents a simplification over the recently proposed two-stage estimator (Bakk et al., 2022) by having only two steps instead of multiple sub-steps in estimating the measurement model.

We discussed model identification of the unconditional model, derived an Expectation Maximization algorithm for efficient estimation of both steps and presented second-step asymptotic standard errors that account for the variability in the first step. The simplified two-step procedure makes it possible to apply the standard theory of Gong and Samaniego (1981) for obtaining unbiased standard errors, a further improvement over the two-stage estimator. An effective initialization strategy, using (dissimilarity–based) cluster analysis, was also proposed.

In the simulation study, we observed that the performance of the proposed estimator in terms of bias is very similar to the benchmark simultaneous (full-information ML) estimator—and similar to that of the two-stage estimator—with nearly no efficiency loss. The two-step estimator was up to 4 times faster than the simultaneous estimator. It should be mentioned that, in conditions where the entropy of the LC model is low, all estimators show relatively higher variability and bias, a finding in line with previous research on stepwise estimators for single-level LC models (Vermunt, 2010) .

In the real data example, we found interesting lower and higher level class configurations, consistent with existing literature on the topic of citizenship norms (see, e.g., (Oser et al., 2022)). In the structural model, the model allows us to investigate the associations between covariates and the latent classes, including the possibility of group-level heterogeneous effects of covariates on lower class membership. In addition, we found a considerable CPU running time difference between the one-step and the two-step estimators, which was even larger than what we observed in the more controlled simulation environment. More specifically, whereas the former required 4.5 min to reach convergence, the latter only needed 22 s. From an applied user's perspective, such a CPU time gain can be substantial on a larger scale. As an example, consider a data set with larger low- and high- level sample sizes: if simultaneous estimation took 2 h, our two-step estimator would produce final estimates in only roughly 12 min. We expect, based on existing literature on two-step estimators (see, e.g., (Di Mari & Maruotti, 2022)), such a gap to increase in model complexity - i.e. number of lower/higher level classes and/or available predictors. The difference in time is also multiplied if the models are estimated repeatedly, for example when different sets of covariates or different numbers of latent classes are explored.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

There are some issues that deserve future research. First, while we describe two possible approaches for class selection in Sect. 3.3, this is not the main focus of the current work. Further research should investigate class selection using the different estimators. Second, we have proposed estimates for the asymptotic variance–covariance matrix based on the outer product of the score. Deriving Hessian– and/or sandwich–based (White, 1982) standard errors, e.g. for small high-level sample size and complex sampling scenarios, can be interesting topics for future work. Third, we have discussed multimodality of the likelihood surface as a long-standing well-known characteristic feature related, in general, to mixture models. The EM algorithm's properties have been largely studied over the years - i.e., monotonicity, and global convergence (see, e.g., (Redner & Walker, 1984)). The EM has several advantages, e.g., low cost per iteration, economy of storage and ease of programming. However, in practice, due to multimodality, convergence to global or local optima depends on the choice of the starting point (Wu, 1983) . As such, there is no systematic, neither theoretical nor simulation based, study of the behavior of the EM with two-step estimators. We speculate that, given that the second step operates in a lower dimensional space compared to simultaneous estimation, two-step estimators should somewhat restrain the initialization problem. This point, being not the focus of the current work, certainly deserves specialized attention. For this, and related matters, we defer to future research.

Acknowledgments

The Authors thank Johan Lyrvall for his valuable comments on the manuscript. Di Mari acknowledges financial support from a University of Catania grant (Starting Grant FIRE, PIACERI 2020/2022), and Oser by a European Union grant (ERC, PRD, project number 101077659).

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A. Appendix A: Computation of the Score Vector for the Multilevel Latent Class Model

A.1. The unconditional multilevel LC (first step)

Let us reparametrize the unconditional multilevel LC model of Equation (2) according to the following log-linear equations

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

$$\begin{cases} \log\left[\dfrac{\phi_{h|t}}{1-\phi_{h|t}}\right] = \beta_{h|t} \\ \log\left[\dfrac{\omega_m}{\omega_1}\right] = \alpha_m \\ \log\left[\dfrac{\pi_{t|m}}{\pi_{1|m}}\right] = \gamma_{t|m}, \end{cases} \tag{23}$$

In addition, let us conveniently rewrite (9) as follows

$$Q(\boldsymbol{\alpha}, \boldsymbol{\Gamma}, \boldsymbol{B}) = Q(\boldsymbol{\alpha}) + Q(\boldsymbol{\Gamma}) + Q(\boldsymbol{B}), \tag{24}$$

where $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_M)'$, $\boldsymbol{\Gamma}$ is a $T-1 \times M$ matrix with elements $\gamma_{t|m}$, for $m = 1, \ldots, M$ and $t = 2, \ldots, T$, $\boldsymbol{B}$ is an $H \times T$ matrix with elements $\beta_{h|t}$ for $t = 1, \ldots, T$ and $h = 1, \ldots, H$, and

$$Q(\boldsymbol{\alpha}) = \sum_{j=1}^{J}\sum_{m=1}^{M} \hat{u}_{j,m} \log(\omega_m) \tag{25}$$

$$Q(\boldsymbol{\Gamma}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} \hat{v}_{i,j,t,m} \log(\pi_{t|m}) \tag{26}$$

$$Q(\boldsymbol{B}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} \hat{v}_{i,j,t,m}\{Y_{ijh} \log(\phi_{h|t}) + [1 - Y_{ijh}] \log(1-\phi_{h|t})\}. \tag{27}$$

Recalling that $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}' = \partial Q/\partial \boldsymbol{\theta}'$, the ij-th contribution to the score has the following three blocks, with generic elements

$$\mathbf{s}_{ij,\boldsymbol{\theta}_1}(\hat{\alpha}_m) = \hat{u}_{j,m} - \omega_m \tag{28}$$

$$\mathbf{s}_{ij,\boldsymbol{\theta}_1}(\hat{\gamma}_{t|m}) = (\hat{q}_{i,j,t|m} - \pi_{t|m})\hat{u}_{j,m} \tag{29}$$

$$\mathbf{s}_{ij,\boldsymbol{\theta}_1}(\hat{\beta}_{h|t}) = \sum_{m=1}^{M} \hat{v}_{i,j,t,m}(Y_{ijh} - \phi_{h|t}). \tag{30}$$

Thus, an estimate of $\boldsymbol{\Sigma}_{11}$ can be obtained as follows

$$\hat{\boldsymbol{\Sigma}}_{11} = N^{-1} \sum_{j=1}^{J}\sum_{i=1}^{n_j} \mathbf{s}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{B}})\, \mathbf{s}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{B}})' \tag{31}$$

A.2. The multilevel LC model with covariates (second step)

Let us define $\pi^{ij}_{t|m} = \frac{\exp(\boldsymbol{\gamma}'_{tm}\mathbf{Z}_{ij})}{1+\sum_{s=2}^{T}\exp(\boldsymbol{\gamma}'_{tm}\mathbf{Z}_{ij})}$. The Q function of Equation (17) can be rewritten under the log-linear parametrizations introduced above, except for the second block which is as follows

$$Q(\boldsymbol{\Gamma}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{m=1}^{M}\sum_{t=1}^{T} \hat{v}_{i,j,t,m} \log(\pi^{ij}_{t|m}) \tag{32}$$

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

The second block of the ij-th contribution to the score as generic $K+1$ contributions

$$\mathbf{s}_{ij,\boldsymbol{\theta}_2}(\hat{\boldsymbol{\gamma}}_{tm}) = \hat{u}_{j,m}(\hat{q}_{i,j,t|m} - \pi^{ij}_{t|m})\mathbf{z}_{ij}. \tag{33}$$

B. Extra Tables and Figures

TABLE 8.
Average relative efficiency for the two-step and two-stage estimator relative to the one-step estimator (SD over benchmark one-step SD), averaged over covariate effects.

Condition	Two-stage	Two-step
1	0.985	0.985
2	1.000	1.000
3	0.99	0.99
4	0.995	0.995
5	0.989	0.989
6	0.998	0.998
7	0.997	0.997
8	1.000	1.000
9	0.997	0.997
10	0.999	0.999
11	0.999	0.999
12	1.000	1.000
13	1.000	1.000
14	1.000	1.000
15	1.000	1.000
16	1.000	1.000
17	1.000	1.000
18	1.000	1.000
19	0.983	0.983
20	0.998	0.998
21	0.993	0.993
22	1.005	1.005
23	0.996	0.996
24	1.004	1.004
25	1.001	1.001
26	0.999	0.999
27	0.997	0.997
28	1.000	1.000
29	1.001	1.001
30	0.999	0.999
31	0.999	0.999
32	1.000	1.000
33	1.000	1.000
34	1.000	1.000
35	1.000	1.000
36	1.000	1.000

References

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology, 30*(1), 27–80.

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics, 37*(6A), 3099–3132.
Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling, 21*(3), 329–341.
Bakk, Z., Di Mari, R., Oser, J., & Kuha, J. (2022). Two-stage multilevel latent class analysis with covariates in the presence of direct effects. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(2), 267–277.
Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika, 83*, 871–892.
Bartolucci, F., Montanari, G. E., & Pandolfi, S. (2015). Three-step estimation of latent Markov models with covariates. *Computational Statistics & Data Analysis, 83*, 287–301.
Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis, 41*(3–4), 561–575.
Catania, L., & Di Mari, R. (2021). Hierarchical Markov-switching models for multivariate integer-valued time-series. *Journal of Econometrics, 221*(1), 118–137.
Catania, L., Di Mari, R., & Santucci de Magistris, P. (2022). Dynamic discrete mixtures for high-frequency prices. *Journal of Business & Economic Statistics, 40*(2), 559–577.
Da Costa, L. P., & Dias, J. G. (2015). What do Europeans believe to be the causes of poverty? A multilevel analysis of heterogeneity within and between countries. *Social Indicators Research, 122*(1), 1–20.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.
Di Mari, R. , Bakk, Z. , Oser, J. , & Kuha, J. (2022). Multilevel latent class analysis with covariates: Analysis of cross-national citizenship norms with a two-stage approach. Under review.
Di Mari, R., Bakk, Z., & Punzo, A. (2020). A random-covariate approach for distal outcome prediction with latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(3), 351–368.
Di Mari, R., & Maruotti, A. (2022). A two-step estimator for generalized linear models for longitudinal data with time-varying measurement error. *Advances in Data Analysis and Classification, 16*, 273–300.
Di Zio, M., Guarnera, U., & Rocci, R. (2007). A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics & Data Analysis, 51*(5), 2573–2585.
Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software, 40*(8), 1–18.
Fagginger Auer, M. F., Hickendorff, M., Van Putten, C. M., Bèguin, A. A., & Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education, 29*, 144–159.
Gassiat, É., Cleynen, A., & Robin, S. (2016). Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing, 26*(1–2), 61–71.
Gnaldi, M., Bacci, S., & Bartolucci, F. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification, 10*, 53–70.
Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics, 9*, 861–869.
Goodman, L. A. (1974). 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*, 1179–1259.
Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models* (Vol. 1). Cambridge University Press.
Grilli, L., Marino, M. F., Paccagnella, O., & Rampichini, C. (2022). Multiple imputation and selection of ordinal level 2 predictors in multilevel models: An analysis of the relationship between student ratings and teacher practices and attitudes. *Statistical Modelling, 22*(3), 221–238.
Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting timss and pirls combined data: multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics, 10*(4), 2405–2426.
Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 7*(4), 121.
Hagenaars, J. A. (1990). *Categorical longitudinal data - Loglinear analysis of panel, trend and cohort data*. Sage.
Hooghe, M., & Oser, J. (2015). The rise of engaged citizenship: The evolution of citizenship norms among adolescents in 21 countries between 1999 and 2009. *International Journal of Comparative Sociology, 56*(1), 29–52.
Hooghe, M., Oser, J., & Marien, S. (2016). A comparative analysis of "good citizenship': A latent class analysis of adolescents' citizenship norms in 38 countries. *International Political Science Review, 37*(1), 115–129.
Horn, M. L. V., Fagan, A. A., Jaki, T., Brown, E. C., Hawkins, J. D., Arthur, M. W., & Catalano, R. F. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research, 43*(2), 289–326.
Huang, G. H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika, 69*(1), 5–32.
Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. In H. M. H. Lu & H. Luu (Eds.), *KDD: Techniques and applications* (pp. 21–34). World Scientific.
Köhler, H., Weber, S., Brese, F., Schulz, W., & Carstens, R. (2018). *ICCS 2016 user guide for the international database: IEA International Civic and Citizenship Education Study 2016*. Amsterdam: The International Association for the Evaluation of Educational Achievement (IEA).

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press

Lukociene, O., Varriale, R., & Vermunt, J. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology, 40*(1), 247–283.
Lyrvall, J. , Di Mari, R. , Bakk, Z. , Oser, J. , & Kuha, J. (2023). multilevlca: An r package for single-level and multilevel latent class analysis with covariates. arXiv preprint arXiv:2305.07276.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 281–297). University of California Press.
Magidson, J. (1981). Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research, 10*, 177–194.
Magidson, J., & Vermunt, J. (2004). Latent class models Latent class models. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 175–198). Sage.
Maruotti, A., & Punzo, A. (2021). Initialization of hidden markov and semi-markov models: A critical evaluation of several strategies. *International Statistical Review, 89*(3), 447–480.
Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(2), 180–197.
McCutcheon, A. L. (1987). *Latent Class Analysis*. Sage.
Morselli, D., & Glaeser, S. (2018). Economic conditions and social trust climates in Europe over ten years: An ecological analysis of change. *Journal of Trust Research, 8*(1), 68–86.
Mutz, R., & Daniel, H. (2013). University and student segmentation: Multilevel latent-class analysis of students' attitudes towards research methods and statistics. *British Journal of Educational Psychology, 83*(2), 280–304.
Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*(2), 479–482.
Oberski, D. L., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling, 20*, 409–428.
Oser, J., Di Mari, R., & Bakk, Z. (2023). Data preparation for citizenship norm analysis, international association for the evaluation of educational achievement (IEA) 1999–2009-2016. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/AKS42
Oser, J., & Hooghe, M. (2013). The evolution of citizenship norms among s candinavian adolescents, 1999–2009. *Scandinavian Political Studies, 36*(4), 320–346.
Oser, J., Hooghe, M., Bakk, Z., & Di Mari, R. (2022). Changing citizenship norms among adolescents, 1999–2009-2016: A two-step latent class approach with measurement equivalence testing. *Quality & Quantity, 2022*, 1–19.
Ouyang, J., & Xu, G. (2022). Identifiability of latent class models with covariates. *Psychometrika, 87*(4), 1343–1360.
Paccagnella, O., & Varriale, R. (2013). Asset Ownership of the Elderly Across Europe: A Multilevel Latent Class Analysis to Segment Countries and Households. In N. Torelli, F. Pesarin, & A. Bar-Hen (Eds.), *Advances in Theoretical and Applied Statistics* (pp. 383–393). Berlin, Heidelberg: Springer, Berlin Heidelberg.
Park, J., & Yu, H. T. (2018). Recommendations on the sample sizes for multilevel latent class models. *Educational and Psychological Measurement, 78*(5), 737–761.
Parke, W. R. (1986). Pseudo maximum likelihood estimation: the asymptotic distribution. *The Annals of Statistics, 14*, 355–357.
Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review, 26*(2), 195–239.
Rindskopf, D. (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*(2), 441–462.
Ruelens, A., & Nicaise, I. (2020). Investigating a typology of trust orientations towards national and European institutions: A person-centered approach. *Social Science Research, 87*, 102414.
Skrondal, A., & Kuha, J. (2012). Improved regression calibration. *Psychometrika, 77*(4), 649–669.
Tomczyk, S., Hanewinkel, R., & Isensee, B. (2015). Multiple substance use patterns in adolescents: A multilevel latent class analysis. *Drug and Alcohol Dependence, 155*, 208–214.
Venables, W. N. , Smith, D. M. , & the R Core Team. (2013). An introduction to R. notes on R: A programming environment for data analysis and graphics version 3.0.0. http://cran.r-project.org/doc/manuals/R-intro.pdf
Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*(1), 213–239.
Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research, 17*(1), 33–51.
Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*, 450–469.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society, 50*(1), 1–25.
Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics, 11*, 95–103.
Zhang, X., van der Lans, I., & Dagevos, H. (2012). Impacts of fast food and the food retail environment on overweight and obesity in China: A multilevel latent class cluster approach. *Public Health Nutrition, 15*(1), 88–96.

Manuscript Received: 26 JUL 2022
Published Online Date: 6 AUG 2023

https://doi.org/10.1007/s11336-023-09929-2 Published online by Cambridge University Press