University of Catania

Doctoral Program in Economics, Management and Decision Making

XXXVII Cycle

Complex latent variable modeling for multivariate hierarchical data

by

Johan Lyrvall

Supervisor:
Professor Roberto Di Mari
Department of Economics and Business
University of Catania

Co-supervisor:
Professor Francesco Drago
Department of Economics and Business
University of Catania

Complex latent variable modeling for multivariate hierarchical data

by
Johan Lyrvall

A thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics, Management and Decision Making

Abstract

This thesis extends, in four independent chapters, the stepwise multilevel latent class analysis with covariates methodological framework. In recent decades, the latent class analysis methodological scholarship has developed sophisticated methodologies, including innovative estimation approaches and open-source software implementations, yet many of these methodological advancements have not yet been extended to the multilevel modeling context. With the increasing availability of hierarchical data, such as cross-national surveys, multilevel latent class analysis is becoming increasingly more applicable, and the utility of extending state-of-the-art methods increasingly greater.

The first chapter, co-authored by Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser, and Jouni Kuha, makes two contributions. First, it compiles state-of-the-art methodologies for multilevel latent class analysis with covariates, describing benchmark model specifications and estimation approaches, and detailing initialization issues and model selection alternatives. Second, it proposes the novel `R` package `multilevLCA` package as the first open-source software implementing the described state-of-the-art methodologies, and the first software overall automatically implementing some stepwise and sequential routines. As such, it contributes both to the methodological statistics literature and to the open-source statistical software literature. This chapter is published as the article *Multilevel Latent Class Analysis: State-of-the-Art Methodologies and Their Implementation in the R Package multilevLCA* in the journal *Multivariate Behavioral Research* [1].

The second chapter, co-authored with Zsuzsa Bakk, Jennifer Oser, and Roberto Di Mari, proposes a novel bias-adjusted three-step estimation approach for multilevel latent class analysis models with covariates. This contributes, to the methodological statistics literature, with the first modeling option for multilevel latent class analysis models with covariates involving the practically appealing property of allowing the applied researcher to work with an explicit dependent variable. This chapter is published as the article *Bias-Adjusted Three-Step Multilevel Latent Class Modeling with Covariates* in the journal *Structural Equation Modeling: A Multidisciplinary Journal*[2].

The third chapter, co-authored with Jouni Kuha and Jennifer Oser, contributes with a novel two-step estimation approach for multilevel latent class analysis models with covariates and non-equivalence of measurement, to the methodological statistics literature. This approach accounts for violations of a standard modeling assumption and extends benchmark estimation routines to modeling contexts which are common in applied cross-national survey research. This chapter is published as the article *Two-Step Multilevel Latent Class Analysis in the Presence of Measurement Non-Equivalence* in the journal *Structural Equation Modeling: A Multidisciplinary Journal* [3].

The fourth chapter, co-authored with Roberto Di Mari and Jouni Kuha, extends benchmark estimation routines for multilevel latent class analysis models with covariates to the Bayesian statistical framework. The novel Bayesian routines contribute to the methodological statistics literature with an alternative methodology with improved performance relative to the standard frequentist approach in common problematic modeling contexts involving small samples. This chapter is work in progress on a more extensive study.

[1] `https://doi.org/10.1080/00273171.2025.2473935`
[2] `https://doi.org/10.1080/10705511.2023.2300087`
[3] `https://doi.org/10.1080/10705511.2025.2490946`

Summary

In four independent chapters, this thesis extends the stepwise latent class analysis methodological framework to the multilevel modeling context. With the increasing availability of hierarchical data like cross-national survey data, the utility of extending state-of-the-art methods to the multilevel modeling context is becoming increasingly greater.

The first chapter provides an overview of recommended methodologies for multilevel latent class analysis, and proposes an innovative software solution for their implementation: the `R` package `multilevLCA`. This more theoretical contribution provides the most comprehensive compilation of benchmark model assumptions and modeling strategies yet; and this more practical contribution provides the first open-source software solution to implement many of the discussed recommended methodologies, and the first to implement all of them in the same package. Specifically, the contribution focuses on standard model assumptions in multilevel latent class analysis, model selection, and estimation approaches. The discussed estimation approaches include the one-step approach and the two existing stepwise approaches for multilevel latent class analysis prior to the development of the chapter, namely, the two-stage approach and the benchmark two-step approach. Toward the end of the chapter, the recommended methodologies compiled in the theoretical overview and implemented in `multilevLCA` are illustrated by means of a real-data analysis of citizenship norms.

The second chapter develops a novel estimation approach for multilevel latent class models: a bias-adjusted three-step approach. The development of this approach involves an extension of the equivalent, original approach for the single-level modeling context. In the single-level context, the bias-adjusted three-step estimator is known to have somewhat less attractive statistical properties, with respect to bias and efficiency, than the two-step estimator; this second chapter shows that these relative properties are retained in the multilevel context. Nevertheless, the multilevel extension of the method retains the practical property of allowing the applied researcher to work with an explicit dependent variable meant to represent the latent class structure, which is considered greatly advantageous by many applied researchers. The bias and efficiency of the method are evaluated on the basis of a simulation study, with comparisons to the one-step approach and the two-step approach. Toward the end of the chapter, the proposed bias-adjusted three-step approach for multilevel models is illustrated by means of a real-data analysis of citizenship norms.

The third chapter extends the two-step estimation approach for multilevel latent class models to the measurement non-equivalence modeling context. For example, measurement non-equivalence is relevant in multilevel research involving cross-national surveying in which differences such as survey language and political contexts may confound the functioning of the survey items as measurements of the latent class structure. This contribution is the first extension of the benchmark two-step approach to the measurement non-equivalence modeling context, and the first extension of any stepwise estimation approach for multilevel latent class models to the measurement non-equivalence modeling context. By means of a simulation study, the method is shown to have comparable statistical properties, with respect to bias and efficiency, to conventional two-step analysis in the absence of measurement non-equivalence, and to single-level latent class analysis in the absence of measurement non-equivalence. Toward the end of the chapter, the proposed two-step approach for multilevel models with measurement non-equivalence is illustrated by means of a real-data analysis of citizenship norms.

The fourth chapter extends the two-step estimation approach in single-level latent class analysis and multilevel latent class analysis to the Bayesian statistical framework. The contributions of the

previous three chapters are put forth within the more conventional frequentist statistical framework. Bayesian estimation routines are known to typically perform better than frequentist estimation routines when the sample size is small. On the basis of a simulation study, the proposed Bayesian approach is shown to have this more advantageous statistical property relative to the conventional, frequentist two-step approach, with respect to bias and efficiency, while retaining the broad pattern of statistical properties of the frequentist approach. Toward the end of the chapter, the proposed Bayesian two-step approach is illustrated by means of two real-data analyses, for the single-level and multilevel modeling contexts, of jazz artists' professional recognition and cultural participation.

Contents

Chapter 1

Multilevel latent class analysis: State-of-the-art methodologies and their implementation in the `R` package `multilevLCA`

Johan Lyrvall, Roberto Di Mari[1]*, Zsuzsa Bakk*[2]*, Jennifer Oser*[3]*, & Jouni Kuha*[4]

1.1 Introduction

Latent class (LC) analysis (Goodman, 1974a; Lazarsfeld & Henry, 1968; McCutcheon, 1979) is used to classify units into discrete types based on a set of observed categorical variables. The clustering is modeled as an underlying discrete variable with some number of categories or *latent classes*. LC analysis has been applied in diverse research domains in the social sciences and beyond. For example, in political research, Oser (2022) identified repertoires of political participation; in educational research, Hickendorff, van Putten, Verhelst, and Heiser (2010) identified patterns of mental strategies for division problems among elementary school students; in substance use research, Bray, Watson, Salisbury-Afshar, Taylor, and McGuire (2023) identified types of opioid users among patients in the emergency department.

A basic assumption of standard LC analysis is that the units of analysis are independent of each other. This conditional independence assumption is often violated when the data have a multilevel, or hierarchical structure, for example when we observe voters within countries, students within schools, or patients within hospitals. In hierarchical data, units within groups are likely to be systematically more similar than units across groups.

To account for the higher-level dependencies in the hierarchical data, the baseline LC model can be extended by modeling a second categorical LC variable at the higher (group) level. In such a *multilevel LC model*, the distribution of the lower-level classes is allowed to vary between the higher-level classes. This random effect is effectively nonparametric (Aitkin, 1999; Finch & French, 2014; Laird, 1978; Vermunt, 2003), thus avoiding strict distributional assumptions. For instance, in their multilevel LC analysis of financial product ownership across European countries, Bijmolt, Paas, and Vermunt (2004) identified 14 individual-level consumer segments and found that the prevalence of these segments varied between 7 country-level clusters. For example, the consumer segment that

[1]Department of Economics and Business, University of Catania, Italy
[2]Department of Methodology and Statistics, Leiden University, Netherlands
[3]Department of Politics and Government, Ben-Gurion University of the Negev, Israel
[4]Department of Statistics, London School of Economics, United Kingdom

was the largest in the cluster of countries in North-Central Europe was rather small in the cluster of countries in North-Western Europe.

In LC analysis, identifying the clustering structure of the data is usually only the first step of the empirical investigation. The research interest usually lies in the relationship between the classes and some covariates, or predictors. In the multilevel LC model, covariates can be included both on the lower level and on the higher level. For instance, in their multilevel LC analysis of adolescent smoking behavior across communities, Henry and Muthén (2010) first identified three individual-level clusters - heavy smokers, moderate smokers, and nonsmokers, and two community-level clusters - low-use communities and high-use communities. Subsequently, they analyzed the regression relationship between smoking behavior and lower-level covariates such as school performance and academic aspirations, and the regression relationship between community type and higher-level covariates such as the proportion of youth living in poverty.

Historically, multilevel LC models were estimated using the traditional *one-step approach*, which involves fitting the full model simultaneously (Lazarsfeld & Henry, 1968; Vermunt, 2003). While the one-step approach has attractive statistical propoperties - when the LC model is correctly specified, it is efficient and asymptotically unbiased - it also comes with serious defects (see e.g. the discussion in Bakk & Kuha, 2018). Whenever covariates are added or removed, the whole model needs to be refitted and the effective definitions of the latent classes can change. This complicates model interpretation and model selection. Furthermore, the one-step approach does not fit with the logic of most applied researchers, who tend to view the regression model as a distinct component that should be estimated only after the clustering model has been built. Therefore, the general recommendation is to use *stepwise estimation approaches* (Asparouhov & Muthén, 2014). These were traditionally only available in single-level LC analysis, but recent methodological advancements have shown how they can be extended to multilevel LC models (Bakk, Di Mari, Oser, & Kuha, 2022; Di Mari, Bakk, Oser, & Kuha, 2023b; Lyrvall, Bakk, Oser, & Di Mari, 2024).

Stepwise approaches avoid the defects of the one-step approach by separating the estimation of the measurement model from the subsequent estimation of the structural model. Among the available stepwise approaches, the *two-step approach* is known to be the most efficient, least biased, most direct, and most flexible option (Bakk & Kuha, 2018; Di Mari et al., 2023b). The *two-stage approach* (Bakk et al., 2022) is slightly less direct but otherwise largely shares the same properties as the two-step approach. Compared to the one-step approach, the two-step and two-stage approaches come with enhanced algorithmic stability and improved speed of convergence (Di Mari et al., 2023b). Regardless of which estimation approach is applied, the number of classes on the higher level and the lower level is taken as given. Because the complexity of the underlying clustering structure in the data tends to be unknown a priori, identification of the optimal number of classes is typically the first step of applied LC analysis.

In light of these recent methodological contributions, the first aim of this article is to provide a compilation of state-of-the-art methods for multilevel LC analysis with covariates. We describe benchmark model specifications and estimation approaches. In addition, we detail initialization issues and model selection alternatives. Targeting both beginning LC analysts and more advanced LC analysts, we hope to strike a satisfying balance between user-friendly ground-up exposition and technical detail.

A lack of general and easily available software solutions has limited the dissemination of these estimation and model selection approaches in the applied multilevel LC analysis literature. The recently published `R` package `multilevLCA` (Di Mari & Lyrvall, 2024) was developed to fill this gap. The package is available from the Comprehensive `R` Archive Network at

`http://cran.r-project.org/package=multilevLCA`.

The second aim of this article is to propose the `multilevLCA` package to the open-source statistical software literature. While the functionalities discussed in this paper can be implemented in specialized software like Latent GOLD (Vermunt & Magidson, 2021b) and Mplus (Muthén & Muthén, 2017), these software options are commercial and offer fewer automatic implementations of stepwise and sequential routines. In this paper we focus on open-source software. We present the capabilities and syntax of `multilevLCA`. The presentation is organized in the article alongside the corresponding

LC analysis methodological exposition, to closely connect software implementation with theory. The software contribution has been written in such a way that we hope that this article can serve as a stand-alone reference for application of `multilevLCA`.

The `multilevLCA` package is both the first freeware-software to implement stepwise estimation of multilevel LC models with covariates and the first to estimate multilevel LC models with both dichotomous and polytomous indicators. `multilevLCA` has the most comprehensive set of model specifications and estimation approaches; estimating single- and multilevel LC models, with and without covariates, using the one-step, two-stage, and two-step approaches. The semi-automatic implementation of model selection in the package is more straightforward and efficient compared to when each model of interest needs to be fitted separately, which is the case when using other freeware-software for LC analysis.

The only existing freeware-software for multilevel LC analysis with covariates is the `R` package `glca` (Kim, Jeon, Chang, & Chung, 2022), but it is limited to the one-step approach, with no implementation of stepwise approaches. Moreover, it does not have the capacity to model polytomous indicators, which are typically used in applied research. As such, the scope of the use of `glca` is somewhat limited compared to `multilevLCA`. The comprehensive functionalities of `multilevLCA` also extend the freeware-software state-of-the-art in single-level LC analysis with covariates. Existing packages for it include the `R` packages `poLCA` (Linzer & Lewis, 2011) and `MultiLCIRT` (Bartolucci, Bacci, & Gnaldi, 2014), but they estimate only single-level models using the one-step approach. The more complete alternative for single-level LC modeling is the `Python` package `StepMix` (Morin et al., 2023), with `R` interface `stepmixr` (Lacourse et al., 2024), which also implements stepwise estimation. However, unlike `multilevLCA`, `StepMix` does not compute maximum-likelihood standard errors of the regression parameters for the covariates, which is the statistical benchmark, instead applying the bootstrap method.

This article offers a comprehensive review of the key aspects of multilevel LC analysis with covariates, and a hands-on guide to the implementation of these techniques using the `multilevLCA` package. In the next section, we present the multilevel LC model and the `multilevLCA` syntax. Then, we describe possible estimation strategies for the model and their implementation in `multilevLCA`, including strategies for class selection and initialization and a benchmark simulation study of performance and estimation times. Next, we illustrate key features of `multilevLCA` by means of an empirical example, and conclude with a summary.

1.2 Model specifications

1.2.1 Theoretical framework

Let Y_{ih} denote the response of unit $i = 1, \ldots, N$ on the categorical item $h = 1, \ldots, H$, with possible values $Y_{ih} = 1, \ldots, R_h$, and let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iH})'$ denote the full response vector for the same unit. The elements of the vector are treated as observed indicators of the categorical latent variable X_i, with possible values $\{1, \ldots, T\}$. The single-level latent class (LC) model defines the unconditional probability of observing a particular response pattern $\mathbf{Y}_i$ as a mixture of T class-specific probabilities, that is,

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X_i = t) P(\mathbf{Y}_i | X_i = t). \tag{1.1}$$

Here, the mixture weight $P(X_i = t)$ describes the unconditional probability that unit i belongs to class t, while the mixture component $P(\mathbf{Y}_i | X_i = t)$ describes the conditional probability of a particular response pattern $\mathbf{Y}_i$ given class t. The responses of the different indicators are assumed to be conditionally independent given class membership (the *local independence assumption*), leading to

$$P(\mathbf{Y}_i) = \sum_{t=1}^{T} P(X_i = t) \prod_{h=1}^{H} P(\mathbf{Y}_{ih}|X_i = t) = \sum_{t=1}^{T} P(X_i = t) \prod_{h=1}^{H} \prod_{r=1}^{R_h} \phi_{rh|t}^{I(Y_{ih}=r)}, \tag{1.2}$$

where the quantity $\phi_{rh|t}$ is the probability of giving response r on item h given class t, and $I(Y_{ih} = r)$ is equal to 1 if unit i gives response r on item h, and 0 otherwise. For ease of notation, we will use $P(\mathbf{Y}_{ih}|X_i = t)$ to denote $\prod_{r=1}^{R_h} \phi_{rh|t}^{I(Y_{ih}=r)}$ in what follows.

Figure 1.1 graphically illustrates the single-level LC model defined in (1.2). The arrows describe a causal relationship from the LC variable X_i to the indicators Y_{ih}. There are no arrows between the indicators, reflecting the local independence assumption.

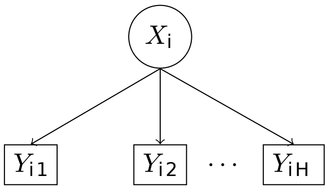

Figure 1.1: The single-level latent class model, with categorical indicators Y and a categorical latent class variable X.

In a multilevel LC model we take the lower-level units $i = 1, \ldots, n_j$ (e.g. individual respondents) to be nested within higher-level units $j = 1, \ldots, J$ (groups, e.g. countries). Let W_j be a higher-level categorical latent variable with possible categories $m = 1, \ldots, M$, and probabilities $P(W_j = m) = \omega_m > 0$, and let X_{ij} now be a lower-level categorical latent variable that is defined conditional on the values of W_j, with possible values $t = 1, \ldots, T$ and conditional probabilities $P(X_{ij} = t|W_j = m) = \pi_{t|m} > 0$. We collect all ω_m and $\pi_{t|m}$ respectively in the M-vector ω, and the $M \times T$ matrix $\mathbf{\Pi}$. The multilevel (random-effect) LC model for $\mathbf{Y}_{ij}$ can be specified as

$$P(\mathbf{Y}_i) = \sum_{m=1}^{M} P(W_j = m) \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right\}, \tag{1.3}$$

where we assume that the conditional response probabilities of items Y_{ijh} depend on higher-level class membership only through X_{ij}. The model specified in (1.3) is similar to the multilevel item response model (Gnaldi, Bacci, & Bartolucci, 2016), but with categorical latent variables on both levels.

While the assumption of conditional independence between Y_{ijh} and W_j given X_{ij} is not necessary for model identification, it is a standard assumption in multilevel LC analysis for enhancing model interpretation (Lukočienė, Varriale, & Vermunt, 2010; Vermunt, 2003). The higher-level LC variable is typically included when it cannot be assumed that the distribution of the lower-level LCs be invariant across higher-level units j (this point is exemplified in a substantive analysis in Section 1.5).

In Figure 1.2, we graphically illustrate the multilevel LC model defined in (1.3). The absence of arrows from the higher-level LC variable W_j to the indicators Y_{ih} reflect their conditional independence given the lower-level LC variable X_{ij}.

Higher-level and lower-level covariates can be included to predict class membership. Let $\mathbf{Z}_{ij} = (1, \mathbf{Z}'_{1j}, \mathbf{Z}'_{2ij})'$ be a vector of K covariates, which can be defined on the higher level ($\mathbf{Z}'_{1j}$) and the lower level ($\mathbf{Z}'_{2ij}$). On the higher level, we consider the following multinomial logistic model

$$P(W_j = m|\mathbf{Z}_j^H) = \frac{\exp(\alpha'_m \mathbf{Z}_j^H)}{1 + \sum_{l=2}^{M} \exp(\alpha'_l \mathbf{Z}_j^H)}, \tag{1.4}$$

where $\mathbf{Z}_j^H = (1, \mathbf{Z}'_{1j})'$, and α_m are regression coefficients for $m = 2, \ldots, M$. When only the intercept term is included, then α_m is equal to the log-odds $\log(\omega_m/\omega_1)$.

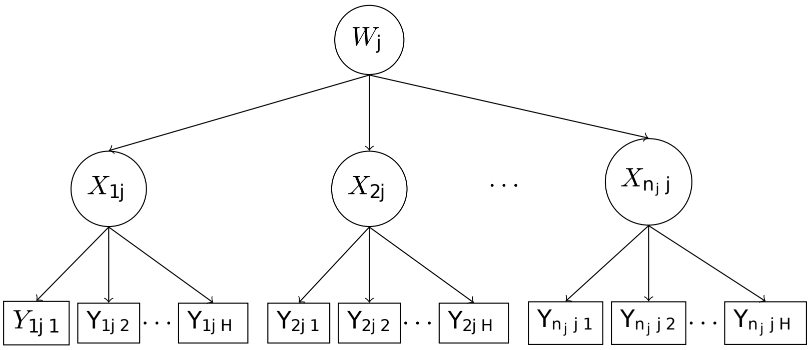

Figure 1.2: The multilevel latent class model, with categorical indicators Y, a categorical lower-level latent class variable X, and a categorical higher-level latent class variable W.

On the lower level, class membership probabilities can be parameterized in the following analogous way,

$$P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) = \frac{\exp(\gamma'_{tm}\mathbf{Z}_{ij})}{1 + \sum_{s=2}^{T} \exp(\gamma'_{sm}\mathbf{Z}_{ij})}, \tag{1.5}$$

where γ_{tm} is a vector of regression coefficients for each $t = 2, \dots, T$, and $m = 1, \dots, M$. When only the intercept term is included, so that $\mathbf{Z}_{ij} = 1$, then γ_{tm} is equal to the log-odds $\log(\pi_{t|m}/\pi_{1|m})$. As can be seen, this parametrization allows the effects of $\mathbf{Z}_{ij}$ on X_{ij} to vary across different m. The methodological exposition throughout this article holds also for the equivalent constrained parametrization in which the slopes are held fixed across different m and only the intercepts are allowed to vary (Di Mari et al., 2023b; Vermunt, 2005). For generality of exposition, we focus on the unconstrained parametrization without fixed slopes in (1.5).

We further assume that the indicators Y_{ijh} are conditionally independent from the covariates given lower-level class membership. With these assumptions, the multilevel LC model for $P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij})$ can be written as

$$P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij}) = \sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_j^H) \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right\}. \tag{1.6}$$

The conditional response probabilities $P(Y_{ijh}|X_{ij} = t)$ define the LC *measurement model*, while the conditional class membership probabilities $P(W_j = m|\mathbf{Z}_j^H)$ and $P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij})$ define the LC *structural models*.

The multilevel LC model with covariates defined in (1.6) is graphically illustrated by means of a path diagram in Figure 1.3. The assumption of conditional independence between the indicators and the covariates given lower-level class membership is reflected in the absence of arrows from $\mathbf{Z}_{ij}$ and $\mathbf{Z}_j^H$ to the Y_{ijh}.

As noted above, multilevel LCA is typically applied when the distribution of the lower-level LCs X_{ij} cannot be assumed to be invariant across higher-level units j. The strategy of capturing this invariance by means of a higher-level clustering structure is known as the random-effect approach. This is the approach on which we focus. For completeness, we now briefly describe the alternative fixed-effect approach. In this approach the distribution of X_{ij} is allowed to vary across each of the J higher-level units. This is achieved by treating higher-level unit membership as a (categorical) covariate in a single-level LC model. Let $\mathrm{I}_i^H = (\mathrm{I}_i(1), \dots, \mathrm{I}_i(J))'$ be a collection of vectors $\mathrm{I}_i(j)$ which are equal to unity if i belongs to j and zero otherwise. A fixed-effect multilevel LC model with covariates can be specified as

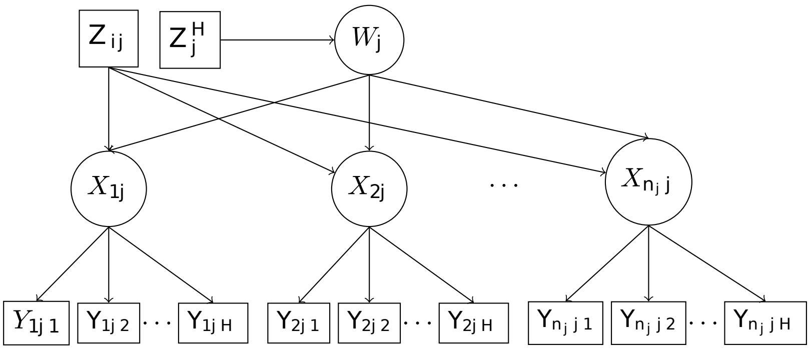

Figure 1.3: The multilevel latent class model with covariates.

$$P(\mathbf{Y}_i|\mathbf{Z}_{ij}) = \sum_{t=1}^{T} P(X_i = t|\mathrm{I}_i^H, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t), \tag{1.7}$$

where, like in the random-effect specification, $P(X_i = t|\mathrm{I}_i^H, \mathbf{Z}_{ij})$ can be parameterized by means of multinomial logistic equations.

1.2.2 Implementation in `multilevLCA`

The syntax used in the `R` package `multilevLCA` is aligned with the notation used in (1.6). The package's multilevel modeling focuses on standard specifications with conditional independence between the items Y_{ijh} and the higher-level LC variable W_j are given the lower-level LC variable X_{ij}. LC models are specified using the function `multiLCA()`, based on some combination of statements about the variables to be included in the model. This is structured by means of the following arguments:

- `data`: Matrix or data frame containing the observed data
- `Y`: Names of `data` columns with indicators
- `iT`: Number of lower-level classes
- `id_high`: Name of `data` column with higher-level id
- `iM`: Number of higher-level classes
- `Z`: Names of `data` columns with covariates in the model for the lower-level classes
- `Zh`: Names of `data` columns with covariates in the model for the higher-level classes

The multilevel LC model with covariates on the higher level and the lower level includes all the variables corresponding to these statements - the indicators $\mathbf{Y}$, specified by `Y`; the lower-level LC variable $X = 1, \ldots, T$, specified by `iT`; the higher-level LC variable $W = 1, \ldots, M$, specified by `id_high` and `iM`; the covariates in the model for the lower-level classes $\mathbf{Z}$, specified by `Z`; and the covariates in the model for the higher-level classes $\mathbf{Z}^H$, specified by `Zh`. The syntax for specifying this model is[5]

```
multiLCA(data, Y, iT, id_high, iM, Z, Zh)
```

[5] `multilevLCA` also estimates multilevel LC models in which the slopes for the lower-level structural model are held fixed across the higher-level classes. This constraint is managed by means of the argument `fixedslopes` in the `multiLCA()` function. The specification `fixedslopes = TRUE` fixes the slopes in the lower-level structural model. The default specification `fixedslopes = FALSE` estimates models without these constraints, which is the focus of this article.

Single-level LC models with covariates and multilevel fixed-effect LC models can be estimated by omitting to specify `id_high`, `iM`, and `Zh` (which default to `NULL`). More specifically, multilevel fixed-effect LC models can be estimated by specifying `Z` as the column name which in random-effect modeling is specified for `id_high`. We illustrate the `multiLCA()` syntax in greater detail by means of real-data examples in Section 1.5.

The next section describes the currently existing approaches for estimating (1.6).

1.3 Methodology

1.3.1 Theoretical framework

Let $\mathbf{Y}_j = (\mathbf{Y}_{1j}, \ldots, \mathbf{Y}_{n_j j})'$ denote the full set of item responses for all lower-level units belonging to higher-level unit j[6]. Let $\theta = (\theta_1', \theta_2')'$ denote the full set of model parameters in (1.6), where θ_1' contains the measurement parameters $\phi_{rh|t}$, and θ_2' contains the structural parameters α_m and γ_{tm}.

Figure 1.4 graphically illustrates the measurement parameters θ_1' by red arrows, and the structural parameters θ_2' by blue arrows.

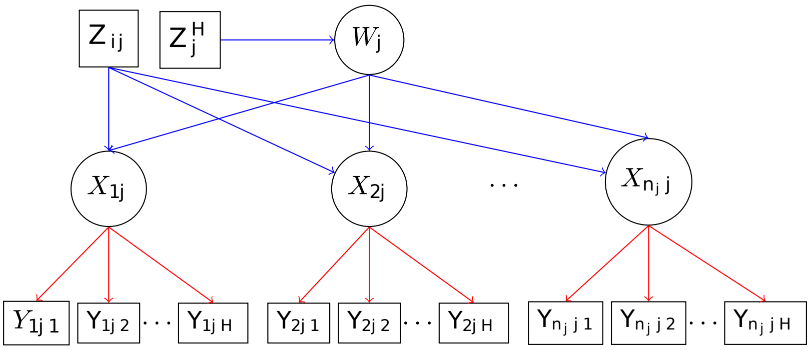

Figure 1.4: The measurement model (red) and structural models (blue).

Maximum-likelihood estimates $\widetilde{\theta}$ can be obtained by maximizing the observed-data log-likelihood function

$$\ell(\theta) = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m | \mathbf{Z}_j^H) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t | W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t) \right\} \right]. \tag{1.8}$$

This is the classical *one-step* approach (Lazarsfeld & Henry, 1968; Vermunt, 2003). It is efficient and asymptotically unbiased when the LC model is correctly specified. However, simultaneous estimation of the measurement model and structural models has serious disadvantages when the correct specification is not known a priori (see e.g. the discussion in Bakk & Kuha, 2018). Whenever the structural model is changed - for example adding or removing covariates - the measurement model will be affected, which distorts the class definitions. In practice, this problem can occur

[6]By default, `multilevLCA` discards any rows with missing values on the items, or incomplete item-response patterns, before estimation. An alternative strategy involves including incomplete item-response patterns by means of full-information maximum-likelihood (FIML) estimation, only discarding any rows with missing values on all the items. The choice between these strategies is managed by means of the argument `incomplete` in the function `multiLCA()`. The default specification `incomplete = FALSE` implements row-wise deletion of incomplete item-response patterns. The alternative specification `incomplete = TRUE` implements the FIML strategy, including incomplete item-response patterns (except fully missing item-response patterns). Regardless of strategy for handling missing values, if covariates are included in the model, rows with missing values in the covariates are removed only in the estimation of the structural part of the LC model, i.e. (see below) step 2 in the two-step estimator, stage 2 in the two-stage estimator, or the single step in the one-step estimator.

to an extent that makes comparisons of estimated models meaningless. As such, the one-step approach complicates model interpretation and model selection. Moreover, simultaneous estimation of complex models involves demanding computations, which renders the one-step approach the more time consuming modeling option for multilevel LC analysis with covariates (Di Mari et al., 2023b).

Stepwise methods overcome the drawbacks of the one-step approach by separating the estimation of the measurement model and structural model. The first stepwise method that was proposed in multilevel LC modeling with covariates is the *two-stage* approach (Bakk et al., 2022; Di Mari, Bakk, Oser, & Kuha, 2023a). Its first stage involves estimating the measurement parameters. This is further broken down into three sub-steps. In the first sub-step, the single-level LC model without covariates is estimated, ignoring the hierarchical structure of the data, by maximizing the log-likelihood function

$$\ell_{\text{stage1.1}}(\theta_1) = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X_{ij} = t) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right], \tag{1.9}$$

where $N = \sum_{j=1}^{H} n_j$, to obtain measurement estimates $\widetilde{\theta}_1$. In the second sub-step, the multilevel LC model without covariates is estimated, keeping the measurement parameters θ_1 fixed at their values from sub-step 1, by maximizing the log-likelihood function

$$\begin{aligned} &\ell_{\text{stage1.2}}(\theta_2|\theta_1 = \widetilde{\theta}_1) \\ &= \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t, \theta_1 = \widetilde{\theta}_1) \right\} \right], \end{aligned} \tag{1.10}$$

where the structural parameters θ_2 now contain only the intercept terms, to obtain structural estimates $\widetilde{\theta}_2$. In the third sub-step, to stabilize the measurement estimates, the multilevel LC model is estimated again, this time keeping the structural parameters θ_2 fixed at their values from sub-step 2, by maximizing the log-likelihood function

$$\begin{aligned} &\ell_{\text{stage1.3}}(\theta_1|\theta_2 = \widetilde{\theta}_2) \\ &= \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|\theta_2 = \widetilde{\theta}_2) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \theta_2 = \widetilde{\theta}_2) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right\} \right]. \end{aligned} \tag{1.11}$$

Stage 2 of the two-stage approach involves adding the covariates to the multilevel LC model, and estimating the intercept and slope terms θ_2, keeping the measurement parameters fixed at their stage-1 values, by maximizing the log-likelihood function

$$\begin{aligned} &\ell_{\text{stage2}}(\theta_2|\theta_1 = \widetilde{\theta}_1) \\ &= \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_j^H) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t, \theta_1 = \widetilde{\theta}_1) \right\} \right]. \end{aligned} \tag{1.12}$$

The two-stage approach simplifies model interpretation and improves computation time compared to the one-step method, while demonstrating very similar properties when the model assumptions hold (Bakk et al., 2022). However, a difficulty of this approach is estimating asymptotic standard errors of the structural parameters. In the second stage, conditioning on the measurement parameters as if they were known, rather than estimated with sampling error, yields underestimation of the standard errors. Conditioning on this first-stage variability is complicated due to the multiple sub-steps of the first stage.

To address this difficulty, the more straightforward *two-step* approach (Di Mari et al., 2023b) was developed. It simplifies the estimation of the measurement model by means of a single first step. This involves maximizing the log-likelihood function

$$\ell_{\text{step1}}(\theta_1) = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t | W_j = m) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t) \right\} \right], \quad (1.13)$$

to obtain measurement estimates $\widetilde{\theta}_1$. The second step involves estimating the structural parameters, keeping the measurement parameters fixed at their step-1 values, by maximizing the log-likelihood function for the second step as

$$\ell_{\text{step2}}(\theta_2 | \theta_1 = \widetilde{\theta}_1) = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m | \mathbf{Z}_j^H) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t | W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t, \theta_1 = \widetilde{\theta}_1) \right\} \right]. \quad (1.14)$$

The two-step approach retains the attractive properties of the two-stage method, with the additional benefits of easy-to-derive asymptotic standard errors, and even greater computational efficiency (Di Mari et al., 2023b).

The estimation approaches that were presented in this section take the number of classes on the higher level, M, and the lower level, T, as given. Selecting these values is a distinct but equally fundamental task. In Section 1.4, two model selection approaches are described.

1.3.2 Implementation in `multilevLCA`

Because of its attractive properties, the two-step approach is the default estimator in the `R` package `multilevLCA`. Users can also choose to estimate LC models using the one-step and two-stage approaches. This makes `multilevLCA` the first `R` package, and the first freeware software in any programming language, to implement stepwise estimation of multilevel LC models with covariates.

Estimation approaches are managed using the argument `fixedpars` in the function `multiLCA()`. One-step, two-stage, and two-step estimation of the multilevel LC model with covariates on the higher level and the lower level are implemented by means of the syntax

```
# One-step estimation:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, fixedpars = 0)

# Two-stage estimation:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, fixedpars = 2)

# Two-step estimation (the default):
multiLCA(data, Y, iT, id_high, iM, Z, Zh, fixedpars = 1)

# Equivalent two-step estimation:
multiLCA(data, Y, iT, id_high, iM, Z, Zh)
```

The estimators are labeled by the total number of fixed parameters; in one-step estimation, no parameters are kept fixed (`fixedpars = 0`); in two-stage estimation, the fixed parameters are obtained from two consecutive sub-steps (`fixedpars = 2`); in two-step estimation, the fixed parameters are obtained from a single step (`fixedpars = 1`).

Regardless of which estimator is used, estimation is performed using the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). When covariates are included, the M step of the EM algorithm uses a Newton-Raphson (NR) algorithm. For computational efficiency, the EM and NR algorithms are implemented by integration of `C++` code (Eddelbuettel & François, 2011; Eddelbuettel & Sanderson, 2014).

In stepwise estimation, the starting values for the EM algorithm are particularly important because subsequent steps are conditional on estimates from previous steps. `multilevLCA` implements an initialization strategy based on Di Mari et al. (2023b).

For the measurement model, the initialization strategy involves the following hierarchical procedure:

1. Fit a single-level LC model with T classes to the pooled data $(\mathbf{Y}_{11}, \ldots, \mathbf{Y}_{n_J J})$, ignoring the multilevel structure. To initialize the class proportions $P(X_i = t)$, perform a k-modes clustering on the dummy-coded data, with $k = T$. Use the relative sizes of the resulting clusters for the initialization. From the single-level class solution, retain the estimates for the conditional response probabilities $P(Y_{ijh}|X_{ij} = t)$, and the modal posterior class assignments[7] $\widetilde{X}_{ij}$. The estimates for $P(Y_{ijh}|X_{ij} = t)$ are passed to the EM-algorithm as starting values.

   For computational speed and stability, the class proportions $P(X_i = t)$ can be initialized by the following alternative strategy. First, perform a principal component analysis on the dummy-coded data. Retain the first principal components that together explain at least 85% of the total variance, or retain the first half of all principal components, if this is a greater number. Second, perform a k-means clustering on the reduced data, with $k = T$. Use the relative sizes of the resulting clusters for the initialization.

2. Compute the relative sizes of $\widetilde{X}_{ij}$ within each higher-level unit j. On the resulting $J \times T$ table, perform a k-means clustering, with $k = M$. Let $\widetilde{W}_j$ be the resulting clusters. The relative sizes of $\widetilde{W}_j$ are passed to the EM-algorithm as starting values for the higher-level class proportions $P(W_j = m)$.

In the function `multiLCA()`, the choice between the k-modes strategy and the k-means on principal components strategy is managed using the logical argument `kmea`. The default argument is `kmea = TRUE`, which indicates the k-means on principal components strategy. The user also has the option to specify custom starting values. This can be done by specifying, in the `multiLCA()` call, the argument `startval` (which defaults to `NULL`) as the name of the `data` column containing starting values for the lower-level class membership of each lower-level unit. The three initialization strategies are implemented by means of the syntax

```
# k-means on principal components initialization:
multiLCA(data, Y, iT, id_high, iM, Z, Zh)

# k-modes initialization:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, kmea = FALSE)

# user-specified starting values:
multiLCA(data, Y, iT, id_high, iM, Z, Zh, startval)
```

For the structural model, the initialization strategy is used to handle label switching on the higher level. Keeping the conditional response probabilities fixed cannot prevent that higher-level class labels can be switched, as there are still $M!$ equivalent permutations of them. This is handled by initializing the intercept in α_m and the intercept in γ_{tm} at the measurement model estimates for $\log(\omega_m/\omega_1)$ and $\log(\pi_{t|m}/\pi_{1|m})$, respectively, while initializing the slope parameters in α_m and the slope parameters in γ_{tm} at zero.

[7] The modal posterior class assignment is the class for which the posterior class membership probability $P(X_{ij} = t|\mathbf{Y}_{ij})$, which describes the probability of belonging to class t given the observed response pattern $\mathbf{Y}_{ij}$, is the greatest. Using the Bayes rule (Goodman, 1974a, 1974b; Hagenaars, 1992; MacLahlan & Peel, 2000), this quantity can be computed as

$$P(X_{ij} = t|\mathbf{Y}_{ij}) = \frac{P(X_{ij} = t)P(\mathbf{Y}_{ij}|X_{ij} = t)}{P(\mathbf{Y}_{ij})} \tag{1.15}$$

1.4 Model selection

1.4.1 Theoretical framework

The general recommendation in LC analysis with covariates is to perform model selection on the model without covariates, defined in (1.3), and then estimate the full model given this value (Masyn, 2017). In multilevel LC analysis, different approaches can be used to identify the locally optimal number of higher-level classes, M, and lower-level classes, T, among a set of specifications. Using the straightforward *simultaneous* approach, all crossed combinations of the values of interest for M and T are estimated.

Using the generally recommended *sequential* approach (Lukočienė et al., 2010), the optimal values for M and T are selected in a stepwise procedure. First, single-level LC models, defined in (1.2), are estimated to select the optimal number of lower-level classes, T^*. Second, multilevel LC models are estimated, keeping the number of lower-level classes fixed at the step-1 value T^*, to select the optimal number of higher-level classes, M^*. Third, multilevel LC models are estimated again, this time keeping the number of higher-level classes fixed at the step-2 value M^*, to re-select the number of lower-level classes.

The optimal model can be selected based on standard information criteria, such as the Bayesian information criterion (BIC) or the Akaike information criterion (AIC). BIC can be evaluated on the higher level and the lower level separately (e.g. Lukočienė et al., 2010). Another information criterion is the BIC-type approximation of the integrated complete likelihood (ICL-BIC; e.g. Morgan, 2015), which can be defined on the higher level and the lower level separately, wherein a penalty for class separation is added to the BIC.

1.4.2 Implementation in `multilevLCA`

The `R` package `multilevLCA` implements semi-automatic[8] model selection, for model specifications without covariates, using the simultaneous and sequential approaches. This is done using the same syntax as for standard model estimation, in the function `multiLCA()`, with the number of classes on the higher level and the lower level specified as a range of consecutive integers, and model selection approaches managed using the argument `sequential`. The argument `sequential = TRUE` indicates sequential model selection, and the argument `sequential = FALSE` indicates simultaneous model selection. The sequential approach is the default model selection approach.

Consider, for example, the multilevel LC model with an unknown number of lower-level classes, which is taken to be within the range 1-5, and an unknown number of higher-level classes, taken to be within the range 1-3. The syntax for implementing simultaneous and sequential model selection is

```
# Sequential model selection:
multiLCA(data, Y, iT = 1:5, id_high, iM = 1:3)

# Simultaneous model selection:
multiLCA(data, Y, iT = 1:5, id_high, iM = 1:3, sequential = FALSE)
```

Regardless of which model selection approach is implemented, the function call returns the optimal model, and information criteria for all the estimated models. The information criteria include higher- and lower-level BIC, AIC, and higher- and lower-level ICL-BIC. The optimal model is selected based on BIC; with simultaneous model selection, the lower-level BIC, and with sequential model selection, the lower-level BIC for step 1, the higher-level BIC for step 2, and again the lower-level BIC for step 3. This is illustrated by means of a real-data example in Section 1.5.

[8]Semi-automatic in the sense that the package implements model selection only over the range of specifications which is specified by the user.

1.4.3 Performance and estimation time of model selection

To examine the performance and estimation time for the semi-automatic implementation of the simultaneous and sequential model selection approaches in the `multilevLCA` package, we conduct a simulation study. The population model has twelve binary items Y_{ijh}. For all the lower-level classes, the probability of the most likely response is set to 0.8. We vary the number T of lower-level classes X_{ij} from three to five, and the number M of higher-level classes W_j from two to three. The sample sizes on the lower level and the higher level are 500 and 30, respectively.

In all the simulation conditions, the first lower-level class $X_{ij} = 1$ has high probabilities (0.8) of endorsement for all the items and the last lower-level class $X_{ij} = T$ low probabilities (0.2) of endorsement for all the items.

When the number of lower-level classes is $T = 3$, the second class has high probabilities for the first six items $Y_{ij1}, \ldots, Y_{ij6}$ and low probabilities for the last six items $Y_{ij7}, \ldots, Y_{ij12}$. The lower-level class proportions within the first and second higher-level classes are:

- $P(Xij = 1|Wj = 1) = P(Xij = 3|Wj = 2) = 0.19$
- $P(Xij = 2|Wj = 1) = P(Xij = 2|Wj = 2) = 0.31$
- $P(Xij = 3|Wj = 1) = P(Xij = 1|Wj = 2) = 0.51$
- $P(Xij = t|Wj = 3) = 1/T = 0.33$ for all t, when a third higher-level class is modeled

When the number of lower-level classes is $T = 4$, the second and third classes have high probabilities only for the first and last six items, respectively. The lower-level class proportions within the first and second higher-level classes are:

- $P(Xij = 1|Wj = 1) = P(Xij = 4|Wj = 2) = 0.10$
- $P(Xij = 2|Wj = 1) = P(Xij = 3|Wj = 2) = 0.17$
- $P(Xij = 3|Wj = 1) = P(Xij = 2|Wj = 2) = 0.28$
- $P(Xij = 4|Wj = 1) = P(Xij = 1|Wj = 2) = 0.46$
- $P(Xij = t|Wj = 3) = 1/T = 0.25$ for all t, when a third higher-level class is modeled

When the number of lower-level classes is $T = 5$, the second, third, and fourth classes have high probabilities only on the first, mid, and last four items, respectively. In this context, the lower-level class proportions within the first and second higher-level classes are:

- $P(Xij = 1|Wj = 1) = P(Xij = 5|Wj = 2) = 0.06$
- $P(Xij = 2|Wj = 1) = P(Xij = 4|Wj = 2) = 0.10$
- $P(Xij = 3|Wj = 1) = P(Xij = 3|Wj = 2) = 0.16$
- $P(Xij = 4|Wj = 1) = P(Xij = 2|Wj = 2) = 0.26$
- $P(Xij = 5|Wj = 1) = P(Xij = 1|Wj = 2) = 0.43$
- $P(Xij = t|Wj = 3) = 1/T = 0.20$ for all t, when a third higher-level class is modeled

In all the simulation conditions, model selection is performed over a range of values for T and M. The smallest value for these ranges is one, while we vary the highest values by means of the excess above the true number of classes, considering excesses equal to one or three. For example, with a lower-level excess of three for $T = 3$ and a higher-level excess of one for $T = 2$, we perform model selection over 1-6 lower-level classes and 1-3 higher-level classes.

Table 1.1 summarizes the resulting 24 fully crossed simulation conditions. For each of them, we generate 50 random samples.

Sim. cond.	T	M	T-exc.	M-exc.
1	3	2	1	1
2	4	2	1	1
3	5	2	1	1
4	3	3	1	1
5	4	3	1	1
6	5	3	1	1
7	3	2	3	1
8	4	2	3	1
9	5	2	3	1
10	3	3	3	1
11	4	3	3	1
12	5	3	3	1
13	3	2	1	3
14	4	2	1	3
15	5	2	1	3
16	3	3	1	3
17	4	3	1	3
18	5	3	1	3
19	3	2	3	3
20	4	2	3	3
21	5	2	3	3
22	3	3	3	3
23	4	3	3	3
24	5	3	3	3

Table 1.1: Fully crossed simulation conditions based on the true and excess number of lower-level classes T, and the true and excess number of higher-level classes M.

The sequential model selection approach correctly identified the true number of lower-level and higher-level classes for all the simulation conditions and random samples. The simultaneous approach performed equally well for the lower level, while, for the higher level, it yielded a 50/50 success rate across the random samples for 16 of the 24 simulation conditions. For the other simulation conditions, it yielded a success rate of 47-49/50 across the random samples[9].

Figure 1.5 reports the average estimation time for the sequential and simultaneous model selection approaches across the 24 simulation conditions and 50 replications. As expected, the time cost for both approaches tends to be greater when the range of values for the number of classes is larger on the lower level or the higher level. It can clearly be seen that the sequential approach is consistently faster than the simultaneous approach. The time cost for the sequential approach is less sensitive to the range of values for T or M, so that the time cost difference increases when these ranges increase.

1.5 Empirical example: citizenship norms

To illustrate the functionalities of the `R` package `multilevLCA`, we analyze data from the International Civic and Citizenship Education Study 2016 (Schulz et al., 2018) of the International Association for the Evaluation of Educational Achievement (IEA), which have been used to advance political research on citizenship norms (Hooghe & Oser, 2015; Hooghe, Oser, & Marien, 2016; Oser & Hooghe, 2013; Oser, Hooghe, Bakk, & Di Mari, 2023). For details on data cleaning and recoding, see Oser, Di Mari, and Bakk (2023). These data are contained in `multilevLCA` as the data frame `dataIEA`. We can load the package and the data by executing

```
library(multilevLCA)
data("dataIEA")
```

[9] The success rate was 47/50 for simulation condition 21; 48/50 for simulation condition 5; 49/50 for simulation conditions 1, 4, 7, 8, 16 and 22.

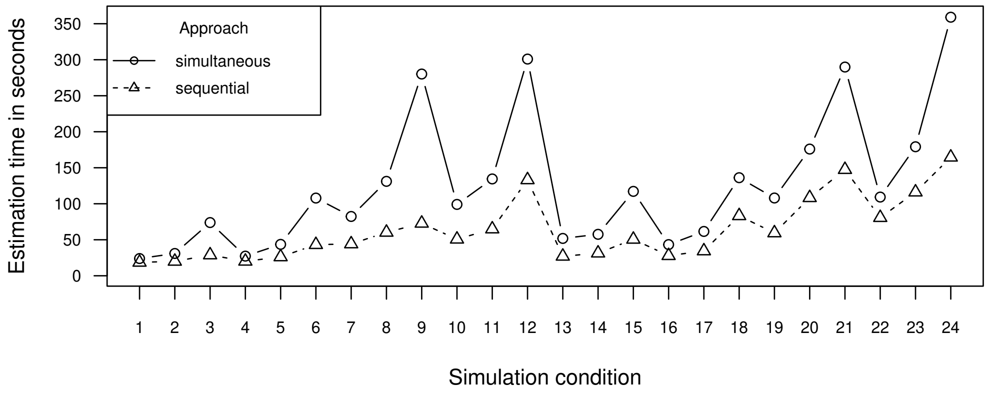

Figure 1.5: Estimation time for the sequential model selection approach and the simultaneous model selection approach, averaged across the 24 simulation conditions and the 50 replications.

We interpret the substantive results in relation to the LC analysis of the same data by Oser, Hooghe, et al. (2023). Prior to their investigation, the political literature on citizenship norms had been focusing on societal-level analyses. The LC analysis informs the literature by taking a person-centered approach and investigating how individuals in different sub-groups of the population adhere to distinct citizenship norms.

As part of a comprehensive evaluation of education systems, the IEA conducted surveys in school classes of 14–year olds to investigate civic education. The use of responses from adolescents to analyze citizenship norms is justified by political research showing that stabilization of individual political attitudes and behaviors occurs rather early in the life cycle (Prior, 2010; Van Deth, Abendschön, & Vollmar, 2011). The survey lists a variety of activities for respondents to rate in terms of importance in order to be considered a good adult citizen. These can be categorized as self-expressive, engaged normative ideals: promoting human rights (*rights*), participating in local activities (*local*), supporting activities to protect the environment (*envir*), participating in peaceful protest (*protest*), and engaging in political conversations (*discuss*); and traditional, duty-based normative ideals: obeying the law (*obey*), working hard (*work*), voting (*vote*), learning about the country's history (*history*), showing respect for government representatives (*respect*), following political news (*news*), and joining a political party (*party*). The answer options "very important" and "quite important" are here coded as 1, while the answer options "not very important" and "not important at all" are coded as 0.

Similar to Oser, Hooghe, et al. (2023), in our LC analysis, we treat the items as observed indicators $\mathbf{Y}_{ij}$ of an underlying structure of citizenship norms X_{ij}, where i denotes a particular student, and j denotes the country in which the school is located. The data contain 90,221 students from 22 countries.

To illustrate the observed response patterns, we print the first three rows below (the observed responses to the questionnaire items are located in columns 5-16).

```
head(dataIEA[,5:16], 3)

obey rights local work envir vote history respect news protest discuss party
   1      1     1    1     1    1       1       1    1       1       0     0
   1      1     1    1     1    1       1       1    1       1       0     0
   1      1     1    1     1    1       1       1    1       0       0     0
```

We begin the illustrative analysis with the five-class single-level LC model without covariates, which was defined in (1.2), replicating the analysis of Oser, Hooghe, et al. (2023), by executing

```
set.seed(2023)
multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5)

CLASS PROPORTIONS:

P(C1) 0.3956
P(C2) 0.3509
P(C3) 0.1111
P(C4) 0.1147
P(C5) 0.0277

RESPONSE PROBABILITIES:

                 C1     C2     C3     C4     C5
P(obey|C)    0.9801 0.9742 0.6335 0.9594 0.3408
P(rights|C)  0.9802 0.9601 0.7386 0.2999 0.0485
P(local|C)   0.9678 0.9079 0.7267 0.3517 0.0527
P(work|C)    0.9364 0.8894 0.5991 0.8532 0.3150
P(envir|C)   0.9800 0.9767 0.7135 0.4771 0.1241
P(vote|C)    0.9727 0.7893 0.6644 0.7476 0.1605
P(history|C) 0.9399 0.8361 0.5992 0.7031 0.1744
P(respect|C) 0.9384 0.8569 0.5357 0.8351 0.1465
P(news|C)    0.9621 0.7171 0.5150 0.7015 0.0783
P(protest|C) 0.8713 0.5701 0.6315 0.1672 0.0516
P(discuss|C) 0.8400 0.1782 0.3945 0.1797 0.0122
P(party|C)   0.6071 0.1439 0.3071 0.1519 0.0177

---------------------------

MODEL AND CLASSIFICATION STATISTICS:

ClassErr       0.1966
EntR-sqr       0.6181
```

At the bottom of the partial `multiLCA()` output above, we can see class separation statistics for the class solution, namely, the average proportion of classification error (`ClassErr`; see Vermunt & Magidson, 2021b), and the entropy-based R^2 (`EntR-sqr`; see Magidson, 1981). To interpret these statistics, consider the task of predicting class membership based on the model parameters (using the modal assignment rule). Based on the average proportion of classification error, we can expect 20% of the respondents to be assigned to the wrong class. Based on the entropy-based R^2, we can expect a 62% improvement of the class prediction when using the response probabilities and class proportions, compared to the prediction using only the class proportions.

The results show that estimated 11.1% and 11.5% of the respondents belong to class 3 and class 4, respectively. Class 3 is corresponding to the "Engaged" class and class 4 to the "Duty" class in Oser, Hooghe, et al. (2023). The youth belonging to class 3 have consistently high conditional probabilities to score 1 (i.e., indicate high importance) on the self-expressive and engaged notions of good citizenship, and consider the traditional and duty-based items to be less important. Class 4 places high importance on the traditional items, except for joining a political party, while placing relatively low importance on the self-expressive items. From a theoretical perspective, the capacity of LCA to identify these two distinctive citizenship norms allows us to address longstanding questions in the literature regarding the socio-demographic characteristics of people who adhere to these different norms.

We can automatically plot the estimated response probabilities by executing

```
plot(out)
```

The resulting plot is shown in Figure 1.6.

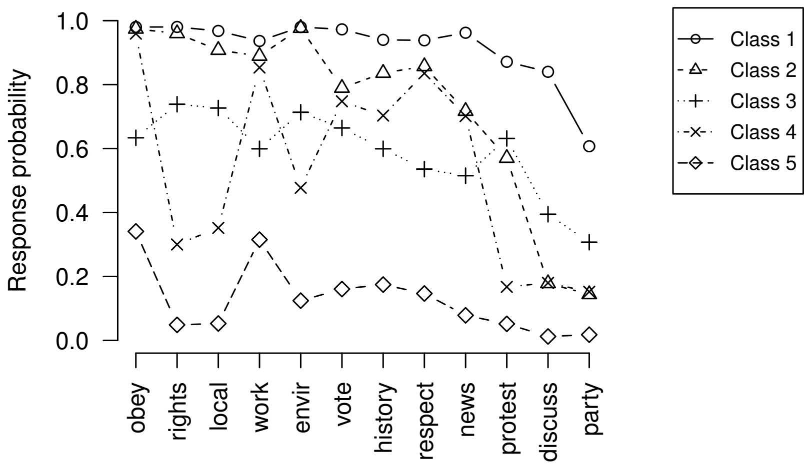

Figure 1.6: Plot generated using the function `multiLCA()`.

To investigate whether the proportion of classification error differs between the classes, we request extensive `multiLCA()` output using the specification `extout = TRUE`. The quantities of interest are contained in the element `mClassErrProb`, which we display below, rounded to two decimal points. The rows of the matrix correspond to true class membership, while columns correspond to predicted class membership. As shown, the expected proportion of correct classification for class 3 (Engaged) and class 4 (Duty) are 73% and 76%, respectively. The youth belonging to class 3 have 9% probability of being assigned to class 4, and those belonging to class 4 a 10% probability of being assigned to class 3.

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               extout = TRUE)
round(out$mClassErrProb, 2)

        C1_pred C2_pred C3_pred C4_pred C5_pred
C1_true    0.87    0.11    0.02    0.01    0.00
C2_true    0.13    0.76    0.07    0.04    0.00
C3_true    0.04    0.13    0.73    0.09    0.01
C4_true    0.01    0.11    0.10    0.76    0.02
C5_true    0.00    0.00    0.04    0.06    0.90
```

The element `mU_modal`, which is returned when `extout = TRUE`, contains the modal class assignment of the units. As shown below, respondents scoring 0 on all the items are estimated to belong to class 5.

```
head(out$mU_modal, 1)

obey rights local work envir vote history respect news protest discuss party
   0      0     0    0     0    0       0       0    0       0       0     0
C1 C2 C3 C4 C5
 0  0  0  0  1
```

Next, we extend the analysis of Oser, Hooghe, et al. (2023) by accounting for the hierarchical structure of the data using the multilevel LC model. The higher-level unit is the country of the respondent (the `dataIEA` column `COUNTRY`). The rationale of this multilevel modeling is that we do not assume the distribution of citizenship norms to be invariant across countries. We could reasonably accept that this distribution would vary across different clusters of countries. We perform model selection on the higher level and, to illustrate how multilevel LC analysis is typically carried out, the lower level. For simplicity of illustration, we consider a small range of values; 1-2 classes on the higher level and 4-5 classes on the lower level (in a more substantive LC analysis of these data, we should reasonably consider larger ranges, such as 1-4 on the higher level and 1-6 on the lower level). In applied LC analysis, the one-class specification is often included in model selection to test for the presence of a clustering structure in the data. We perform model selection using the sequential approach by executing

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 4:5,
               id_high = "COUNTRY", iM = 1:2)

$step1
     BIClow    BIChigh   AIC       ICL_BIClow ICL_BIChigh
iT=4 877289.33 876869.28 876813.64 -          -
iT=5 872987.19 872460.07 872390.24 -          -

$step2
         BIClow    BIChigh   AIC       ICL_BIClow ICL_BIChigh
iT*,iM=1 872987.19 872460.07 872390.24 -          -
iT*,iM=2 869122.92 868554.62 868479.34 952146.46  868554.62

$step3
         BIClow    BIChigh   AIC       ICL_BIClow ICL_BIChigh
iT=4,iM* 873450.73 872997.73 872937.72 942352.88  872997.73
iT=5,iM* 869122.92 868554.62 868479.34 952146.47  868554.62

$optimal
iT= 5
iM= 2
```

The `multiLCA()` output above shows that the model with two higher-level classes and five lower-level classes was selected as the local optimum across the considered specifications. The value $T = 5$ was selected based on the lower-level BIC in the first step, $M = 2$ selected based on the higher-level BIC in the second step, and $T = 5$ re-selected based on the lower-level BIC in the third step.

The function call for model selection returns the results for the optimal model. This is equivalent to directly estimating the model of interest, if it were "known" to be the locally optimal specification, that is, by executing

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               id_high = "COUNTRY", iM = 2)
```

For brevity, we do not print the output for this model. The equivalent fixed-effect model can be estimated by executing

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               Z = "COUNTRY", fixedpars = 0)
```

Again, for brevity, we do not print the resulting output.

Next, we add covariates on both levels, specifying the model defined in (1.6). On the higher level, we consider as covariate the country's gross domestic product (GDP) per capita in constant terms with log transformation (`log_gdp_constant`). These data are obtained from the International Monetary Fund, and included in `dataIEA`. On the lower level, we consider as covariates the respondent's gender (`female`; 1 if the respondent is a girl, 0 if the respondent is a boy) and immigration status of the family (`immigrantfam`; 1 if the respondent comes from a family of immigrants, 0 otherwise). We estimate this model by executing

```
multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
         id_high = "COUNTRY", iM = 2, Z = c("female","immigrantfam"),
         Zh = "log_gdp_constant")

GROUP PROPORTIONS (SAMPLE MEAN):

P(G1) 0.5909
P(G2) 0.4091

CLASS PROPORTIONS (SAMPLE MEAN):

           G1     G2
P(C1|G) 0.2904 0.5494
P(C2|G) 0.4135 0.2729
P(C3|G) 0.1193 0.0884
P(C4|G) 0.1467 0.0667
P(C5|G) 0.0300 0.0226

---------------------------

LOGISTIC MODEL FOR HIGHER-LEVEL CLASS MEMBERSHIP:


MODEL FOR G2 (BASE G1)

                              Alpha   S.E.  Z-score   p-value
alpha(Intercept|G2)          9.2286 0.1748  52.7958 0.0000***
alpha(log_gdp_constant|G2) -0.9376 0.0171 -54.6772 0.0000***

---------------------------

LOGISTIC MODEL FOR LOWER-LEVEL CLASS MEMBERSHIP:


MODEL FOR C4 (BASE C1) GIVEN G1

                            Gamma   S.E.  Z-score   p-value
gamma(Intercept|C4,G1)    -0.7142 0.1052  -6.7866 0.0000***
gamma(female|C4,G1)        0.0789 0.0377   2.0914 0.0365**
gamma(immigrantfam|C4,G1) -0.2994 0.0697  -4.2963 0.0000***

MODEL FOR C4 (BASE C1) GIVEN G2

                            Gamma   S.E.  Z-score   p-value
gamma(Intercept|C4,G2)    -2.0883 0.1337 -15.6217 0.0000***
gamma(female|C4,G2)       -0.0653 0.0499  -1.3093 0.1904
gamma(immigrantfam|C4,G2)  0.4719 0.0867   5.4434 0.0000***
```

```
*** p < 0.01, ** p < 0.05, * p < 0.1
```

As shown in the partial `multiLCA()` output above, the results suggest that 59% of the countries belong to higher-level class 1, while 41% belong to higher-level class 2. The countries belonging to higher-level class 1 emphasize the citizenship norms of lower-level class 2, while the countries belonging to higher-level class 1 emphasize the citizenship norms of lower-level class 1. As such, we can label higher-level class 1 "Mainstream-emphasizing countries", and higher-level class 2 "Maximalist-emphasizing countries". The prevalence of the citizenship norms of lower-level class 4 (Duty) within higher-level class 1 (Mainstream-emphasizing countries) is about twice as high compared to the prevalence within higher-level class 2 (Maximalist-emphasizing countries).

Below the class separation statistics and information criteria, we can see the estimated logistic regression model for higher-level class membership. The negative and highly statistically significant estimate for the effect of GDP per capita - `alpha(log_gdp_constant|G2)` - suggests that wealthier countries have smaller probabilities of belonging to higher-level class 2 relative to higher-level class 1 than less wealthy countries.

Furthermore, we can see the logistic regression parameter estimates for lower-level class membership, conditional on higher-level class membership. For brevity, we comment only on the logistic regression coefficient for gender in the model for membership lower-level class 4 (Duty) relative to lower-level class 1, given higher-level class 1. This coefficient is labeled `gamma(female|C4,G1)` in the above output. The positive sign and statistical significance (at the 5%-level) suggest that, in the countries belonging to higher-level class 1, girls have larger probabilities than boys of belonging to lower-level class 4 relative to lower-level class 1, when controlling for immigration background.

To investigate the posterior class membership probabilities, we specify `extout = TRUE`. We focus on the posterior higher-level class membership probabilities for the countries, which is contained in the element `mPW`, rounding the values to two decimal points (`R` does not display decimal points when the values are very close to 0 or 1). In the printed partial output below, we can see that higher-level class 1 includes, for example, the Nordic countries: Denmark (`DNK`), Finland (`FIN`), Norway (`NOR`) and Sweden (`SWE`). Higher-level class 2 includes, for example, the Asian areas: Hong Kong (`HKG`), South Korea (`KOR`) and Taiwan (`TWN`).

```
out = multiLCA(data = dataIEA, Y = colnames(dataIEA)[5:16], iT = 5,
               id_high = "COUNTRY", iM = 2, Z = c("female","immigrantfam"),
               Zh = "log_gdp_constant", extout = TRUE)
round(out$mPW, 2)

    log_gdp_constant G1 G2
DNK            10.70  1  0
FIN            10.56  1  0
HKG            10.89  0  1
KOR            10.44  0  1
NOR            11.08  1  0
SWE            10.73  1  0
TWN            10.69  0  1
```

1.6 Concluding remarks

We presented the state of the art of multilevel latent class modeling with covariates. The focus was on estimation approaches, model selection, and freeware-software. We presented the theoretical modeling framework, the most advantageous estimation approaches, and recommendations for model selection, including a benchmark simulation study of performance and estimation times for model selection. We gave a tutorial of the user-friendly syntax of the `R` package `multilevLCA` that executes this estimation, visualizes the results, and implements semi-automatic model selection.

The aim of the article was to disseminate the use of advanced multilevel latent class modeling among applied researchers from a variety of academic disciplines. Multilevel latent class analysis has a wide range of applications in fields such as the educational, political, economic, health and behavioral disciplines. There is considerable appeal in this methodology, which allows great flexibility in the parametrization of individual differences in a (possibly multidimensional) phenomenon of interest.

Chapter 2

Bias-adjusted three-step multilevel latent class modeling with covariates

Johan Lyrvall, Zsuzsa Bakk[1]*, Jennifer Oser*[2]*, & Roberto Di Mari*[3]

2.1 Introduction

Latent class analysis is a model-based approach used to create a clustering of units of analysis on the basis of a set of observed indicator variables. The clustering is expressed in terms of a latent variable with some number of discrete categories, or *latent classes*. For example, latent class analysis has been used to identify repertoires of political participation (Oser, 2022), risk profiles for adolescent substance abuse (Lanza & Rhoades, 2013), and patterns of study strategy (Hickendorff et al., 2010). When the data have a multilevel structure, that is, when lower-level units are nested in higher-level units (for example, students nested in a school, patients nested in a centre), the *multilevel* latent class model (Vermunt, 2003) is used to account for higher-level dependencies in the observed variables. Such non-independence arises because respondents in the same group give more similar answers to each other than respondents from different groups.

Multilevel latent class analysis involves introducing a latent variable at the higher level as a categorical random effect defining the distribution of the lower-level clusters within groups. For example, Henry and Muthén (2010) identified a typology of adolescent smokers, non-, moderate, and heavy smokers, and found that the types can be clustered into two community-level segments: low-use and high-use communities. Fagginger Auer, Hickendorff, Van Putten, Béguin, and Heiser (2016) identified four study strategies among students, which could be clustered into four types of teachers with different probabilities of eliciting these study strategies.

Usually, the research interest in latent class analysis lies in explaining latent class membership by covariates, or external variables (for example, neighborhood effects on school performance). Identifying the latent classes (measurement model) is then only the first step of the analysis, after which a more complex model that includes the covariates (structural model) is specified. In most applications, the focus is on the lower-level latent classes and covariates are included only on this level.

In single-level latent class analysis, different approaches for estimation of the covariate effects are available, namely the classical one-step, or simultaneous, approach (Lazarsfeld & Henry, 1968), the two-step approach (Bakk & Kuha, 2018) and the different three-step approaches. The seminal works on three-step LCA were put forth by Vermunt (2010) and Bolck, Croon, and Hagenaars (2004).

[1]Department of Methodology and Statistics, Leiden University, Netherlands
[2]Department of Politics and Government, Ben-Gurion University of the Negev, Israel
[3]Department of Economics and Business, University of Catania, Italy

Extensions include three-step LCA for distal outcomes (Bakk, Tekle, & Vermunt, 2013; Lanza, Tan, & Bray, 2013) and for latent Markov models (Di Mari, Oberski, & Vermunt, 2016). The general recommendation is to use stepwise methods (Asparouhov & Muthén, 2014).

Using the one-step approach, both the measurement and structural model are estimated simultaneously to obtain maximum-likelihood estimates. This approach is apparently natural, however it has serious defects (see e.g. Asparouhov & Muthén, 2014; Bakk & Kuha, 2018; Vermunt, 2010). The full model needs to be re-estimated when a change is made to only one part of it, e.g., increase or decrease in the number of classes, inclusion or exclusion of covariates in the structural model.

Practically, re-estimating the full model can be computationally demanding, especially when the number of potential covariates is large. Furthermore, while the covariates are taken to be class predictors, changes to the structural model can change the latent class solution, especially if underlying model assumptions are violated (Bakk & Kuha, 2018; Masyn, 2017; Vermunt, 2010). This can occur to the extent that comparisons of different estimated structural models become effectively meaningless. In addition, many applied researchers see the introduction of covariates as a separate step after the classification model has been constructed. Different research groups may even build different structural models on the same measurement model.

To avoid the problems of the one-step approach, stepwise methods separate the estimation of the measurement model from the estimation of the structural model. Using the naive three-step approach, (1) the measurement model is estimated alone, (2) units are assigned to latent classes and (3) posterior class assignments are related to the covariates. This was a popular approach as it avoids the defects of the one-step approach, however it introduces a problem of its own: classification error in the second step, yielding bias in the step-3 estimates of the covariate effects (Bolck et al., 2004).

In the last 20 years, many developments have been suggested for single-level latent class analysis to address this issue. The bias-adjusted three-step methods correct the bias by explicitly modeling the classification error in the second step (Bakk et al., 2013; Bolck et al., 2004; Vermunt, 2010). Using the two-step approach (Bakk & Kuha, 2018), (1) the measurement model is estimated alone and (2) the full model is estimated with the measurement parameters held fixed at their step-1 estimates. In the two-step approach, the structural model is estimated conditional on measurement model parameter estimates from step 1, and no actual clustering step is performed.

While the one-step approach can be considered the statistical benchmark in bias and efficiency compared to the stepwise estimators (Bakk & Kuha, 2018; Bakk et al., 2013; Bolck et al., 2004; Vermunt, 2010), applied researchers tend to prefer using the bias-adjusted three step approach, because of the practical appeal of working with an explicit dependent variable. However, while extensions to multilevel latent class models exist for both the one-step approach (Vermunt, 2003) and the two-step approach (Di Mari et al., 2023b), the bias-adjusted three-step approach exists only for single-level latent class models. As such, there is a lack of modeling options for multilevel latent class analysis with covariates with respect to single-level latent class analysis with covariates.

In the current paper, we fill this important gap by introducing a bias-adjusted three-step approach for multilevel latent class analysis. The proposed method is a multilevel extension of the bias-adjusted three-step ML approach for single-level LCA of Vermunt (2010). Our contribution complements the existing set of methodologies for the multilevel context with the typically preferred analytical approach in applied social research. Since the research interest in applied latent class analysis typically lies in the association between covariates and the lower-level latent classification, the proposed method was developed for structural modelling on this level. A similar extension has been developed previously for latent Markov modeling with covariates (Di Mari et al., 2016).

The remainder of the paper is outlined as follows. First we introduce the multilevel latent class model with covariates. Then, we discuss the bias-adjusted three-step ML approach for this model, deriving the correction under standard model assumptions. Subsequently, we report the results of a simulation study in which we compare the method to the one- and two-step methods under different conditions. We next present an empirical application wherein we identify citizenship norms among adolescents and analyze its association with socioeconomic status. The article ends with a summary of the main results and possible directions for future research.

2.2 The multilevel latent class model

Let $\mathbf{Y}_{ij} = (Y_{ij1}, ..., Y_{ijH})$ be a vector of observed responses, where Y_{ijh} denotes the response of low-level unit (individual) $i = 1, \ldots, n_j$ in high-level unit (group) $j = 1, \ldots, J$ on the h-th categorical indicator variable (item) (Vermunt, 2003). For simplicity of exposition, we focus on dichotomous indicators with values 0 and 1. Let $\mathbf{Y}_j = (\mathbf{Y}_{1j}, \ldots, \mathbf{Y}_{n_J j})$ be the set of responses for all low-level units i in high-level unit j. The $\mathbf{Y}_j$ for different j are taken to be independent.

Let W_j be a categorical latent class (LC) variable defined at the higher level, with possible, mutually exclusive values $m = 1, \ldots, M$ and probabilities $\omega_m = P(W_j = m) > 0$. Given a realization of W_j, let X_{ij} be a categorical LC variable defined at the low level with possible, mutually exclusive values $t = 1, \ldots, T$ and conditional probabilities $\pi_{t|m} = P(X_{ij} = t|W_j = m) > 0$. The X_{ij} for the same j are taken to be conditionally independent given W_j, that is,

$$P(X_{ij}, \ldots, X_{n_j j}) = \sum_{m=1}^{M} P(W_j = m) \prod_{i=1}^{n_j} P(X_{ij} = t|W_j = m). \tag{2.1}$$

The simple LC model defines the following probability structure on $\mathbf{Y}_{ij}$,

$$P(\mathbf{Y}_{ij}) = \sum_{t=1}^{T} P(X_{ij} = t) P(\mathbf{Y}_{ij}|X_{ij} = t). \tag{2.2}$$

The probability of observing a particular response pattern $P(\mathbf{Y}_{ij})$ is a linear combination of T class-specific item response probabilities $P(\mathbf{Y}_{ij}|X_{ij} = t)$, where the (unconditional) class proportions $P(X_{ij} = t)$ serve as weights. Furthermore, we make the local independence assumption that the H indicator variables are conditionally independent within the X_{ij}, so that

$$P(\mathbf{Y}_{ij}|X_{ij} = t) = \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) = \prod_{h=1}^{H} \phi_{h|t}^{Y_{ijh}} (1 - \phi_{h|t})^{1-Y_{ijh}}, \tag{2.3}$$

where $\phi_{h|t} = P(Y_{ijh} = 1|X_{ij} = t)$. For simplicity of notation, we express the local independence assumption in terms of the second equality, leading to the following specification of the simple LC model,

$$P(\mathbf{Y}_{ij}) = \sum_{t=1}^{T} P(X_{ij} = t) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t). \tag{2.4}$$

To account for the multilevel structure of the data, the multilevel LC model of (Vermunt, 2003) defines the following probability structure on $\mathbf{Y}_{ij}$,

$$P(\mathbf{Y}_{ij}) = \sum_{m=1}^{M} \omega_m \left\{ \sum_{t=1}^{T} \pi_{t|m} \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t) \right\}. \tag{2.5}$$

The probability of observing a particular response pattern $P(\mathbf{Y}_{ij})$ is a linear combination of M high-level class-specific simple LC models, where the class proportions ω_m serve as weights. This is the so-called non-parametric modeling approach (for the parameteric approach, wherein the higher-level latent variable is continuous, see, e.g., Asparouhov & Muthén, 2014). As can be seen, we make the common assumption that the item response probabilities do not directly depend on W_j, that is, $P(\mathbf{Y}_{ij}|X_{ij} = t, W_j = m) = P(\mathbf{Y}_{ij}|X_{ij} = t)$ (Vermunt, 2003).

The multilevel LC model can be defined in terms of logistic equations. For ω_m we can consider the following nonparametric random-effect model,

$$P(W_j = m) = \frac{\exp(\alpha_{0m})}{1 + \sum_{l=2}^{M} \exp(\alpha_{0l})}, \tag{2.6}$$

while for $\pi_{t|m}$,

$$P(X_{ij} = t | W_j = m) = \frac{\exp(\gamma_{0tm})}{1 + \sum_{s=2}^{T} \exp(\gamma_{0sm})}, \tag{2.7}$$

and for $\phi_{h|t}$,

$$P(Y_{ijh} = 1 | X_{ij} = t) = \frac{\exp(\beta_t^h)}{1 + \exp(\beta_t^h)}, \tag{2.8}$$

where the parameters for the first classes and response categories are set to zero for identification purposes, that is, $\alpha_{01} = \gamma_{01m} = \beta_1^h = 0$. We denote the vector of parameters of the measurement model for the items Y_{ijh} by $\theta_1 = (\phi_{1|1}, \ldots, \phi_{H|T})$. A logistic parametrization of the simple LC model can be obtained similarly on the basis of these equations. For $P(X_{ij} = t)$, the m-subscripts can be omitted from (2.7), while, for $\phi_{h|t}$, (2.8) can be used directly.

Covariates can be included in the multilevel LC model to predict class membership. Let Z_j^H be a high-level covariate, Z_{ij}^L a low-level covariate, and $\mathbf{Z}_{ij} = (Z_j^H, Z_{ij}^L)$. To predict high-level and low-level class membership, we can extend the logistic equations as follows,

$$P(W_j = m | Z_j^H) = \frac{\exp(\alpha_{0m} + \alpha_{1m} Z_j^H)}{1 + \sum_{l=2}^{M} \exp(\alpha_{0l} + \alpha_{1l} Z_j^H)} \tag{2.9}$$

$$P(X_{ij} = t | W_j = m, \mathbf{Z}_{ij}) = \frac{\exp(\gamma_{0tm} + \gamma_{1tm} Z_{ij}^L + \gamma_{2tm} Z_j^H)}{1 + \sum_{s=2}^{T} \exp(\gamma_{0sm} + \gamma_{1sm} Z_{ij}^L + \gamma_{2sm} Z_j^H)}, \tag{2.10}$$

with $\alpha_{11} = \gamma_{11m} = \gamma_{21m} = 0$ for identification. We collect the structural model parameters in the vector by $\gamma = (\gamma_{011}, \ldots, \gamma_{2TM})$. Under these parametrizations we can define the multilevel LC model for $\mathbf{Y}_{ij} | \mathbf{Z}_{ij}$ as

$$P(\mathbf{Y}_{ij} | \mathbf{Z}_{ij}) = \sum_{m=1}^{M} P(W_j = m | Z_j^H) \left\{ \sum_{t=1}^{T} P(X_{ij} = t | W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t) \right\}, \tag{2.11}$$

where we further assume that the observed response patterns $\mathbf{Y}_{ij}$ are conditionally independent of the covariates $\mathbf{Z}_{ij}$ given low-level class membership X_{ij} and high-level class membership W_j. This conditional independence assumption is standard in the multilevel LCA literature (Bakk et al., 2022; Di Mari et al., 2023b). The equivalent conditional independence assumption in single-level LC modelling is standard as well (e.g. Bakk & Kuha, 2018).

In applied multilevel LCA with covariates the research interest typically lies in the lower-level structural model for X_{ij}. Then, covariates Z_j^H may be excluded, or included to control for variation at the higher level, while the substantive research questions regard the effect of Z_{ij}^L on X_{ij}. For example, Bijmolt et al. (2004) fit a multilevel LC model for international consumer segmentation in financial product ownership, excluding predictors of country-level segments and including demographic variables as predictors of individual-level segments.

2.3 Selecting the numbers of latent classes on lower and higher level

The description of the multilevel LC model with covariates above takes the numbers of LCs on lower and higher level as given. In applied LCA, selecting these values is often a distinct exercise. It is generally recommended to carry out this task on the model without the covariates, and then hold the selected numbers of classes fixed when the covariates are added (Masyn, 2017). To do so, two approaches are typically used, namely the sequential and the simultaneous approaches.

The sequential approach of Lukočienė et al. (2010) involves a hierarchical three-step model fitting procedure. First, a set of simple LC models with different numbers of LCs (T) are fitted and the optimal number is selected. Second, this value is held fixed and a set of multilevel LC models with different numbers of high-level LCs (M) are fitted, selecting the best candidate. Third, the selected M is held fixed and model selection is done again at the lower level, to determine the final T.

The more direct simultaneous approach involves estimating the multilevel LC model for all combinations of the different values of T and M of interest. In both model selection approaches, the optimal values of T and M can be selected with standard information criteria, such as the BIC. This can be combined with measures of class separation, such as the entropy-based R^2 (Magidson, 1981), which can be defined at both the lower level and the higher level (Lukočienė et al., 2010).

For a detailed discussion about information criteria and likelihood-based tests for model selection in LC analysis, see Nylund, Asparouhov, and Muthén (2007) (see also Collins, Fidler, Wugalter, & Long, 1993). When the selection of indicators is an issue, see the approaches proposed by Bartolucci, Montanari, and Pandolfi (2016) and by Dean and Raftery (2010).

2.4 Three-step estimation of the multilevel latent class model

In this section we describe a bias-adjusted three-step ML estimation approach for the multilevel LC model, taking the number of LCs on lower level and higher level as given. The procedure involves (i) estimating the simple LC model without covariates, (ii) estimating class membership and classification error, and (iii) estimating a logistic regression model for LC membership while correcting for the classification error introduced in step 2. As such, the proposed method is a multilevel extension of the bias-adjusted three-step ML approach for single-level LCA of Vermunt (2010).

2.4.1 Step 1 - Estimating the multilevel measurement model

In the first step, the LC model without covariates is estimated by maximum likelihood (ML). The model of interest is the multilevel LC model in (2.5), which the researcher may identify by means of model selection, using for example the sequential approach (Lukočienė et al., 2010) or the simultaneous approach.

In step 1 the simple LC model is fitted, ignoring the multilevel structure. Parameter estimates can be obtained by maximizing the following log-likelihood function

$$\ell_{\text{step1}} = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X_{ij} = t) \prod_{h=1}^{H} P(Y_{ijh} | X_{ij} = t) \right], \tag{2.12}$$

We graphically summarize the first step in the left-most panel of Figure 2.1.

2.4.2 Step 2 - Posterior classification and classification error

Given the step-1 parameter estimates, the posterior LC membership probability $P(X_{ij} = t|\mathbf{Y}_{ij})$ can be obtained using Bayes' rule (Goodman, 1974a, 1974b; Hagenaars, 1992; MacLahlan & Peel, 2000) in the following manner

$$P(X_{ij} = t|\mathbf{Y}_{ij}) = \frac{P(X_{ij} = t)P(\mathbf{Y}_{ij}|X_{ij} = t)}{P(\mathbf{Y}_{ij})}. \tag{2.13}$$

From this equation, individual i can be assigned to the LC t on the basis of different classification rules, the most common of which are modal assignment and proportional assignment.

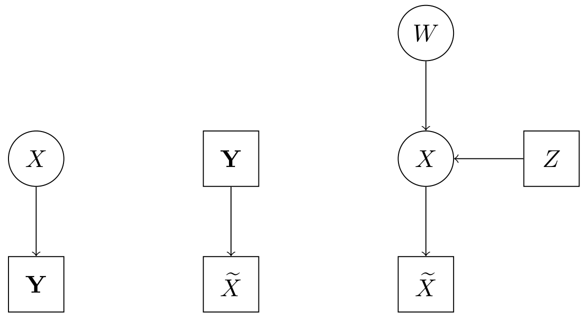

Figure 2.1: The steps of the proposed three-step approach

Let $\widetilde{X}_{ij}$ denote the estimated class membership of individual i in group j. Modal assignment yields a hard partitioning in which i is allocated weight $P(\widetilde{X}_{ij} = s|\mathbf{Y}_{ij}) = 1$ if the posterior membership probability $P(X_{ij} = s|\mathbf{Y}_{ij})$ is the largest, and weight zero otherwise. Proportional assignment yields a soft (crisp) partitioning in which i is allocated weight $P(\widetilde{X}_{ij} = s|\mathbf{Y}_{ij}) = P(X_{ij} = s|\mathbf{Y}_{ij})$. Another classification rule is random assignment, which yields a hard partitioning by assigning i to the $\widetilde{X}_{ij}$ that is randomly drawn from the distribution $P(\widetilde{X}_{ij} = s|\mathbf{Y}_{ij})$ (Goodman, 2007). In most applications the preferred rule is modal assignment because it minimizes classification error (see e.g. Bakk et al., 2013; Vermunt, 2010).

Regardless of which classification rule is used, the assigned class will differ from the true class for some units (Bolck et al., 2004; Hagenaars, 1990). The overall quality of the classification can be quantified by the expected proportion of classification error $P(\widetilde{X}_{ij} = s|X_{ij} = t)$, which expresses the probability of assignment to a certain class conditional on the true class membership. This quantity can be computed as a weighted average over all possible response patterns,

$$\begin{aligned} P(\widetilde{X}_{ij} = s|X_{ij} = t) &= \sum_{\mathbf{Y}} P(\mathbf{Y}|X_{ij} = t) P(\widetilde{X}_{ij} = s|\mathbf{Y}) \\ &= \frac{\sum_{\mathbf{Y}} P(\mathbf{Y}) P(X_{ij} = t|\mathbf{Y}) P(\widetilde{X}_{ij} = s|\mathbf{Y})}{P(X_{ij} = t)}. \end{aligned} \tag{2.14}$$

When the number of unique response patterns is very large, it is convenient to replace the weights $P(\mathbf{Y})$ with their empirical distribution, leading to

$$P(\widetilde{X}_{ij} = s|X_{ij} = t) = \frac{\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j} P(X_{ij} = t|\mathbf{Y}_{ij}) P(\widetilde{X}_{ij} = s|\mathbf{Y}_{ij})}{P(X_{ij} = t)}. \tag{2.15}$$

The second step is summarized graphically in the mid panel of Figure 2.1.

2.4.3 Step 3 - Estimating the multilevel structural model

The starting point of the correction methods for three-step LC analysis is that unadjusted estimation of the relationship between $\widetilde{X}_{ij}$ and $\mathbf{Z}_{ij}$ causes bias toward zero in the covariate effects of interest (Bolck et al., 2004). Given the step-2 LC assignments $\widetilde{X}_{ij}$ and the expected frequency of classification error $P(\widetilde{X}_{ij} = s|X_{ij} = t)$, we can obtain unbiased coefficient estimates defining the relationship between the unobservable LC variable X_{ij} and the covariates $\mathbf{Z}_{ij}$.

Consider the joint distribution of W, X, $\mathbf{Y}$, $\widetilde{X}$ given $\mathbf{Z}$. Given the assumptions of (2.11), that $\mathbf{Y}$ is conditionally independent of $\mathbf{Z}$ and W given X, and that $\widetilde{X}$ only depends on $\mathbf{Y}$, this quantity can be decomposed as

$$P(W, X, \mathbf{Y}, \widetilde{X}|\mathbf{Z}) = P(W|Z^H)P(X|W, \mathbf{Z})P(\mathbf{Y}|X)P(\widetilde{X}|\mathbf{Y}). \tag{2.16}$$

From this equation it is possible to derive the conditional distribution $P(\widetilde{X}|\mathbf{Z})$ by marginalising over lower- and higher-level LCs, and by summing over response patterns, which yields

$$\begin{aligned} P(\widetilde{X}|\mathbf{Z}) &= \sum_W \sum_X \sum_{\mathbf{Y}} P(W|Z^H)P(X|W, \mathbf{Z})P(\mathbf{Y}|X)P(\widetilde{X}|\mathbf{Y}) \\ &= \sum_W P(W|Z^H) \sum_X P(X|W, \mathbf{Z}) \sum_{\mathbf{Y}} P(\mathbf{Y}|X)P(\widetilde{X}|\mathbf{Y}) \\ &= \sum_W P(W|Z^H) \sum_X P(X|W, \mathbf{Z}) \frac{\sum_{\mathbf{Y}} P(X|\mathbf{Y})P(\mathbf{Y})P(\widetilde{X}|\mathbf{Y})}{P(X)} \\ &= \sum_W P(W|Z^H) \sum_X P(X|W, \mathbf{Z})P(\widetilde{X}|X). \end{aligned} \tag{2.17}$$

The first equality follows directly from (2.16), and the second equality re-arranges the summations. The third equality re-writes the last term based on the equation for the expected proportion of classification error. This was defined in the right-hand side of (2.14). The fourth equality simplifies this term into the notation for the expected proportion of classification error. This is the multilevel version of the derivation for the three-step approach in single-level LC modeling (Vermunt, 2010) and latent Markov modeling (Di Mari et al., 2016). As can be seen, the last right-hand side of the last equality of (2.17) is similar to the equation for the multilevel LC model with covariates, but with the step-2 estimates for the expected proportion of classification error in the role of the response probabilities $P(\mathbf{Y}|X)$.

Unbiased estimates of the structural parameters γ and α can be obtained by estimating the right-hand side of (2.17) as a multilevel LC model with covariates, with the low-level class assignments $\widetilde{X}_{ij}$ as a single indicator with known error probabilities $P(\widetilde{X}_{ij}|X_{ij})$. This involves maximizing the log-likelihood function

$$\ell_{\text{step3}} = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W_j = m|Z_j^H) \prod_{i=1}^{n_j} \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij})P(\widetilde{X}_{ij} = s|X_{ij} = t) \right\} \right]. \tag{2.18}$$

This is similar to the second step of the two-step approach, but with the classification error probabilities in the place of the item-specific response probabilities $\prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t)$.

Step 3 is summarized graphically in the right-most panel of Figure 2.1.

2.5 Simulation study

2.5.1 Design

We conduct a simulation study to assess the quality of the proposed bias-adjusted three-step ML approach when the model is correctly specified. As classification rule we use modal assignment, which is the most common in LCA applications and minimizes classification error. The method is compared to the one-step and two-step approaches, which are the currently available modeling options for multilevel latent class analysis with covariates.

We evaluate the relative performance of the three-step estimator based on bias and variation in the estimated covariate effects. The finite-sample quality of stepwise estimators for LCA has been found to depend on class separation and sample size (Bakk & Kuha, 2018; Di Mari et al., 2023b;

Vermunt, 2010). In multilevel LCA, both lower-level and higher-level class separation affect the finite-sample behaviour (Lukočienė et al., 2010).

As population model we use the multilevel LC model with 3 lower-classes X, 2 higher-level classes W, and 10 binary indicator variables Y. Higher-level class proportions are $P(W_j = 1) = 0.6$ and $P(W_j = 2) = 0.4$. We consider one continuous lower-level covariate Z generated from the standard normal distribution. The population slope parameters γ_{1tm} are set to $\gamma_{121} = \gamma_{131} = -0.25$ and $\gamma_{122} = \gamma_{132} = 0.25$. Additionally, we consider population slope parameters equal to zero, $\gamma_{121} = \gamma_{131} = \gamma_{122} = \gamma_{132} = 0$, however, we discuss this in a more limited investigation of statistical power in the results. The first class $X = 1$ has high probability to score 1 on all items, the second class has high probability to score 1 on the last 5 items and 0 on the first 5 items, and class $X = 3$ has high probability to score 0 on all items.

Lower-level class separation is manipulated via the item-response probabilities. In the small, moderate and large separation conditions, the probabilities of the most likely response are 0.7, 0.8 or 0.9, respectively. At the higher level we manipulate class separation via the random intercepts, such that the expected conditional class proportions in the moderate separation condition are equal to $P(X_{ij} = 1|W_j = 1) = 0.29$, $P(X_{ij} = 2|W_j = 1) = 0.33$, $P(X_{ij} = 3|W_j = 1) = 0.38$ and $P(X_{ij} = 1|W_j = 2) = 0.38$, $P(X_{ij} = 2|W_j = 2) = 0.33$, $P(X_{ij} = 3|W_j = 2) = 0.29$, while in the large separation condition they are equal to $P(X_{ij} = 1|W_j = 1) = 0.14$, $P(X_{ij} = 2|W_j = 1) = 0.32$, $P(X_{ij} = 3|W_j = 1) = 0.54$ and $P(X_{ij} = 1|W_j = 2) = 0.60$, $P(X_{ij} = 2|W_j = 2) = 0.25$, $P(X_{ij} = 3|W_j = 2) = 0.15$. For the sample size, we use 100 or 500 lower-level units and 30, 50 or 100 higher-level units.

We quantify class separation as the average entropy-based R^2 (Lukočienė et al., 2010; Magidson, 1981) across the 500 random samples, which is one of the most commonly used measures of class separation in the social sciences. The entropy-based R^2 expresses how much the prediction of class membership improves when using the information on the responses: a value equal to 0 corresponds to no separation and a value equal to 1 corresponds to perfect separation. The average higher-level R^2-entropy for the samples with the moderate and large higher-level separation conditions are 0.76 (0.57 when the lower-level sample size is equal to 100 and 0.96 when the lower-level sample size is equal to 500) and 1.00, respectively. The average lower-level R^2-entropy for the samples with the small, moderate, and large lower-level separation conditions are 0.51, 0.80, and 0.96, respectively.

From each of the 36 crossed simulation conditions (see Table 2.1) we generate 500 random samples. Our simulation setting is similar to previous studies on multilevel LCA (Di Mari et al., 2023b; Lukočienė et al., 2010). Data generation and model estimation is carried out using the R package `multilevLCA` (Lyrvall, Di Mari, Bakk, Oser, & Kuha, 2023)[4].

2.5.2 Results

The results are presented averaged across the four slope parameters γ_{121}, γ_{131}, γ_{122}, and γ_{132} and the 500 replications for the three-, two-, and one-step estimators. Figure 2.2 displays average relative absolute bias and efficiency, measured in terms of the Monte Carlo standard deviation of the relative absolute bias. As can be seen, when the degree of higher-level separation is moderate (simulation conditions 1-18) the three-step method performs essentially identical to the one- and two-step methods. In these conditions the performance varies systematically across the two values for the lower-level sample size: the three estimators are less biased and more efficient when the lower-level sample size is 500 (average R^2_{high}-entropy equal to 0.96) compared to when the lower-level sample size is 100 (average R^2_{high}-entropy equal to 0.57).

The same systematic variation in efficiency is shown for the three estimators when the higher-level separation is large (simulations conditions 19-36, average R^2_{high}-entropy equal to 1.00), however, the performance in terms of bias varies across them in these conditions. While the one-step and two-step methods perform comparably, it can be observed that the performance of the three-step method improves when the degree of lower-level separation is larger. Its performance is problematic when the lower-level separation is small (simulation conditions 19-24, average R^2_{low}-entropy equal to 0.51)

[4] Replication files are available from the corresponding author upon request.

Condition	Sample size		Separation	
	LL	HL	LL	HL
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

Table 2.1: Simulation conditions with crossed combinations of lower-level (LL) and higher-level (HL) sample size and separation for simulation study with 100 replications for each condition

- in these situations the alternative methods are preferred. The bias for the three-step estimator is reduced substantially when the lower-level separation is moderate (simulation conditions 25-30, average R^2_{low}-entropy equal to 0.80) but still slightly higher compared to the one-step estimator and the two-step estimator. The three-step is unbiased when the lower-level separation is large (simulation conditions 31-36, average R^2_{low}-entropy equal to 0.96).

Figure 2.3 displays the ratio of the average SE to SD. For the two-step method we use the corrected SEs to account for measurement uncertainty in its first step (the correction is based on pseudo-ML theory and exploits the full variance matrix; see Bakk & Kuha, 2018; Di Mari et al., 2023b, for details). The one-step SEs and the three-step SEs are based on the expected information matrix. We first note that the one-step SEs are closest to unbiased in most conditions compared to the SEs for the two-step method and the three-step method, while the two-step SEs are over-estimated in most simulation conditions (for the over-correction of the two-step estimator in the single-level context, see Bakk & Kuha, 2018). The three-step SEs are under-estimated, which is in line with previous findings for the single-level context (Bolck et al., 2004; Vermunt, 2010). The downward bias for the three-step method is serious (10-30%) when the higher-level separation is

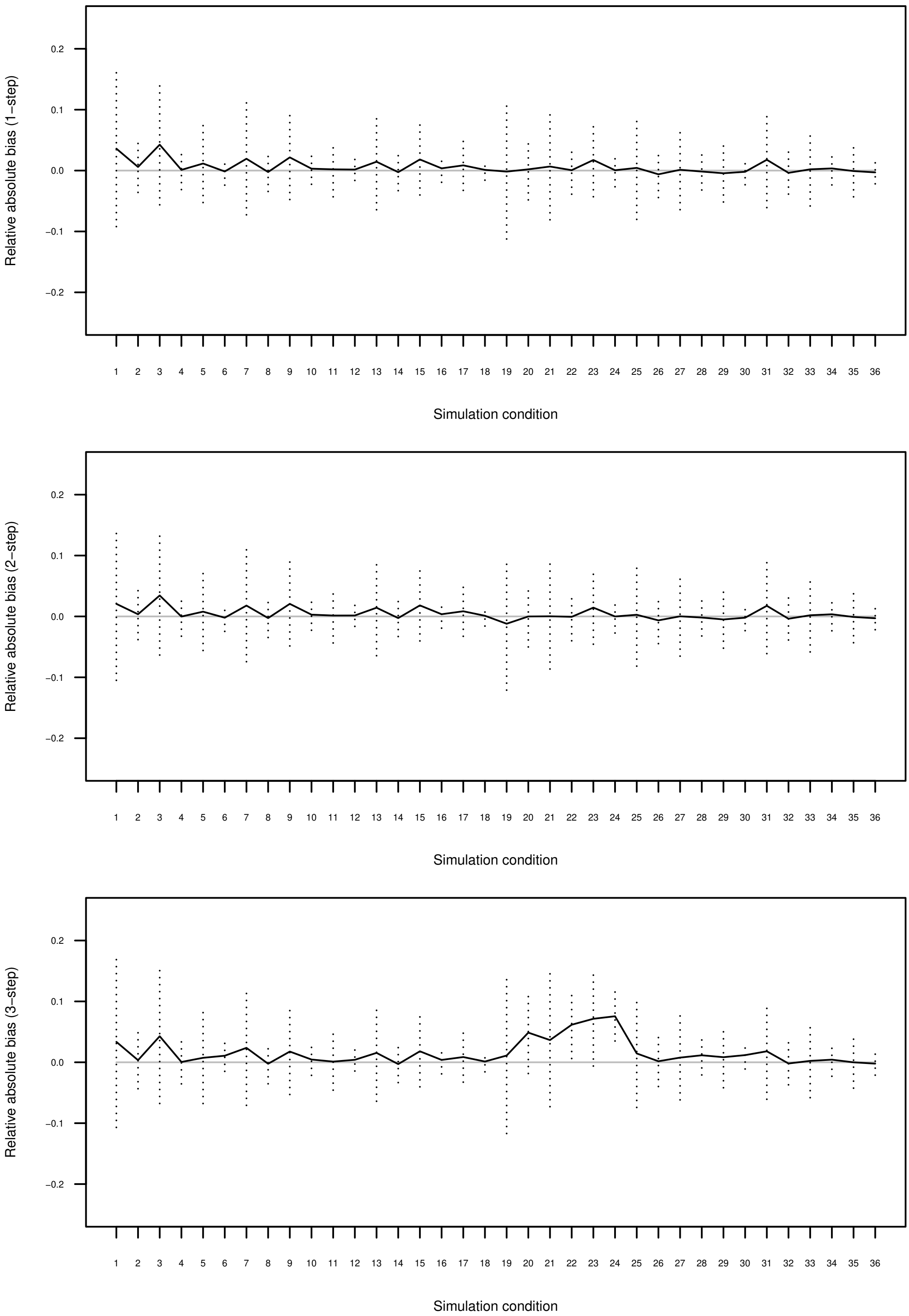

Figure 2.2: Estimated bias (solid) and their Monte Carlo standard deviation (dotted) for the 500 replications per 36 conditions, averaged across the four different γ_{1tm}, separately for the three estimators

moderate and the lower-level separation is small (simulation conditions 1-6), but is reduced as these conditions improve. When the higher-level separation is large and the lower-level separation is large (simulation conditions 31-36), the under-estimation for the three-step method is slight (up to about 5%).

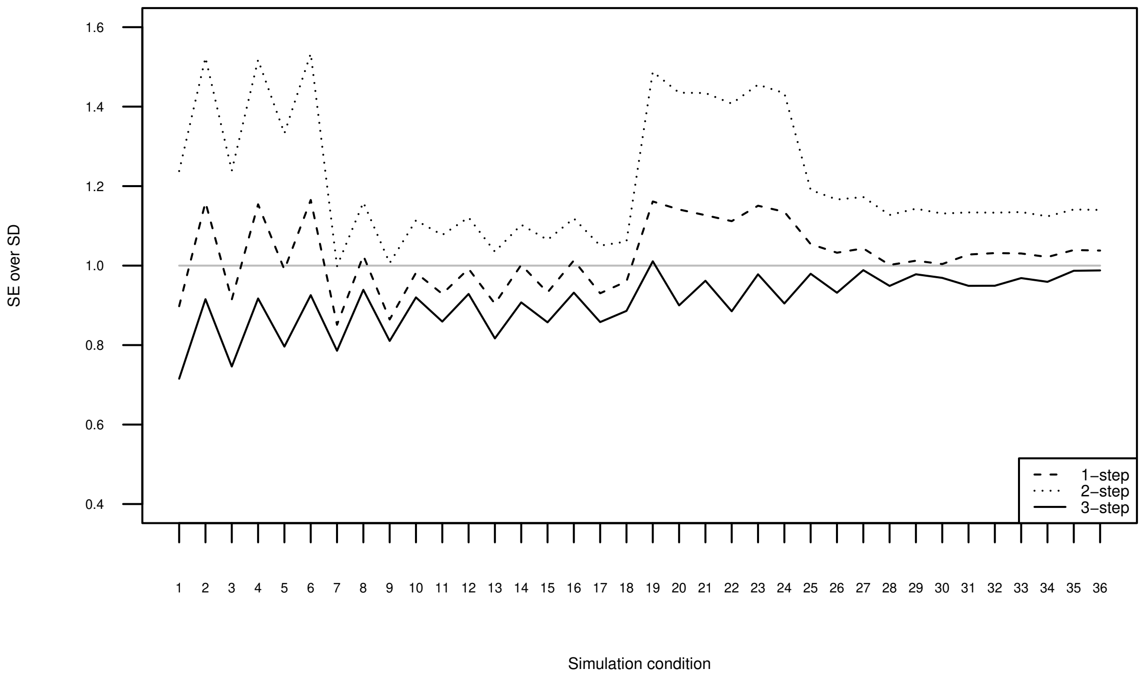

Figure 2.3: Ratio of estimated standard error to Monte Carlo standard deviation for the 500 replications per 36 conditions, averaged across the four different γ_{1tm}

The top panel of Figure 2.4 displays the estimated coverage rates of 95% confidence intervals. The bottom panel displays these values for the samples with population covariate effects equal to zero. The three-step estimator yields undercoverage of the true covariate effects, but this undercoverage is slight (below 5%) when the higher-level separation is large and the lower-level separation is moderate to large (simulation conditions 25-36). When the higher-level separation is large and the lower-level separation is small (simulation conditions 19-24), the performance of the three-step estimator is equally good with zero effect covariates, but problematic with non-zero effect covariates. The problematic performance can be explained by the severe bias in the three-step estimates for the covariate effects in the same conditions. Regardless of the covariate effects, the one-step estimator yields close to correct coverage, while the two-step estimator yields slight overcoverage.

These results are aligned with previous research in single-level LC modeling, in which the bias has been shown to be greater and the efficiency to be lower when the sample size is smaller and the separation is weaker (Bakk & Kuha, 2018; Vermunt, 2010). The present study shows how this performance varies across combinations of separation on the lower level and the higher level in the multilevel LC modeling. How can we explain that the three-step method is sensitive to the lower-level separation when the higher-level separation is large? The explanation lies in the feature that the three-step method involves a classification step wherein the higher-level clustering is ignored, meaning that it does not take into account the information about the hierarchical clustering structure. This loss of information is greater when the higher-level separation is larger. In comparison, the one-step and two-step methods always use the full amount of information on the clustering, making their performance less sensitive to the degree of higher-level separation.

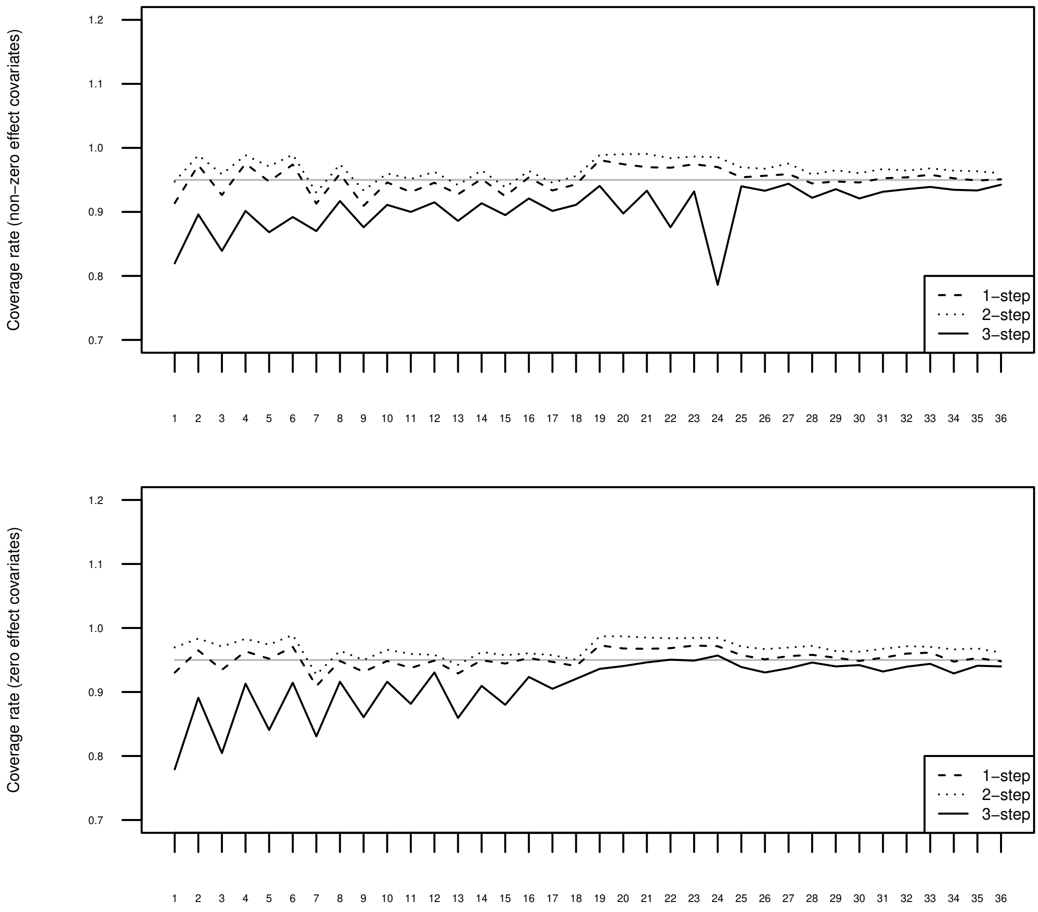

Figure 2.4: Estimated coverage rates of 95% confidence intervals for the 500 replications per 36 conditions, averaged across the four different γ_{1tm}, separately for the simulation study with non-zero covariate effects and the simulation study with zero covariate effects

2.6 An application

To illustrate the proposed bias-adjusted three-step ML method we analyze citizenship norms among adolescents using survey data from the International Civic and Citizenship Education Study, which were collected by the International Association for the Evaluation of Educational Achievement (IEA) in 2016 (Schulz et al., 2018). The survey was conducted in school classes of 14 years old in 22 different areas: Bulgaria, Chile, Colombia, Croatia, Denmark, Dominican Republic, Estonia, Finland, Hong Kong, Italy, Latvia, Lithuania, Malta, Mexico, Netherlands, Norway, Peru, Russia, Slovenia, South Korea, Sweden, and Taiwan. Prior political research has analyzed different waves of the survey to investigate citizenship norms using latent class analysis (Hooghe & Oser, 2015; Hooghe et al., 2016; Oser & Hooghe, 2013; Oser, Hooghe, et al., 2023). The data analysis is carried out using the R package `multilevLCA` (Lyrvall et al., 2023)[5].

Respondents were presented a variety of citizenship norms and asked to rate each item in terms of what defines a good adult citizen. We combine the answer options "very important" and "quite important" into the value 1, while we combine the answer options "not important at all" and "not very important" into the value 0. The 12-item battery includes obeying the law (*obey*), promoting human rights (*rights*), participating in local activities (*local*), working hard (*work*), supporting

[5]Replication files are available from the corresponding author upon request.

activities to protect the environment (*envir*), voting (*vote*), learning about the country's history (*history*), showing respect for government representatives (*respect*), following political news (*news*), participating in peaceful protest (*protest*), engaging in political conversations (*discuss*), and joining a political party (*party*). Table 2.2 shows that the means in the pooled sample range from 0.34 to 0.92.

Item	Mean
obey	0.92
rights	0.84
local	0.82
work	0.86
envir	0.87
vote	0.83
history	0.82
respect	0.83
news	0.77
protest	0.64
discuss	0.46
party	0.34

Table 2.2: Means of the 12 dichotomous items in the pooled sample

Consistent with previous research on the same data (Di Mari et al., 2023b), we consider the model $T = 4, M = 3$. The separation conditions for this model are rather typical in applied LCA (the R^2_{low}-entropy is equal to 0.64 and the average proportion of classification error for step-1 model is equal to 0.17). Compared to the simulation study, the estimated class solution exhibits large class separation at the higher level and somewhere between small and moderate separation on the lower level, which are suboptimal conditions for the performance of the three-step. This allows us to see how severely the performance of the proposed estimator are affected by these suboptimal conditions.

Table 2.3 presents the estimated class solution of this specification. At the lower level, class 1, the "Maximal" adolescent, places importance on all the items; class 2, the "Engaged", scores somewhat lower on all items, placing little importance on engaging in political conversations or joining a political party; class 3, the "Duty", emphasizes the traditional items but not the more self-expressive items; class 4, "Subject", has low probabilities to assign importance to all the items. At the higher level, class 1 (50%) has the highest relative frequency of "Engaged" adolescents, while in class 2 (41%) the majority of youth are "Maximal", and class 3 (9%) has the highest conditional probability for "Duty" students. In all the country-level classes, "Subject" is the least prevalent.

Considering the estimated high-class membership, we can see a geographical pattern. Most of the Nordic-European, Eastern European, and South American countries belong to class 1: Bulgaria, Estonia, Finland, Latvia, Lithuania, Malta, Norway, Slovenia, Sweden, Chile, and Colombia. The countries that belong to class 2 are the Mediterranean-European, North American, Central American, and Asian countries and one South American country: Hong Kong, Russia, South Korea, Taiwan, Dominican Republic, Mexico, Italy, Croatia, and Peru. The Continental Northern European countries belong to class 3: Denmark and the Netherlands.

We now investigate the association between lower-level class membership and maternal education using the proposed bias-adjusted three-step ML approach for multilevel LC models and compare the results to the one- and two-step approaches. The covariate is a binary indicator that takes on the value 1 if the mother has post-high school education level (50% of the respondents), and 0 otherwise (50%). In step 2, respondents are assigned to latent classes by means of modal assignment, since this is the most commonly used assignment rule in applied LCA. Table 2.4 shows that the (lower-level) classes do not differ much in their probability of assignment to the wrong class. The lower-level entropy-based R^2 is equal to 0.64 and the higher-level entropy-based R^2 is equal to 1, which, based on the simulation results, are sub-optimal conditions. On the basis of the simulation results, we can therefore expect a larger difference across the estimated covariate effects between the three-step method and its alternatives than between the one-step method and the two-step method.

	LL Class 1 (Maximal)	LL Class 2 (Engaged)	LL Class 3 (Duty)	LL Class 4 (Subject)
Class proportion				
HL Class 1 (0.5000)	0.3292	0.4676	0.1584	0.0448
HL Class 2 (0.4091)	0.5714	0.2803	0.1190	0.0293
HL Class 3 (0.0909)	0.2064	0.2842	0.4783	0.0312
Response probability				
obey	0.9788	0.9217	0.8793	0.4080
rights	0.9819	0.9456	0.4364	0.1337
local	0.9698	0.9145	0.4243	0.1391
work	0.9368	0.8450	0.7824	0.3820
envir	0.9834	0.9622	0.5251	0.2207
vote	0.9699	0.7477	0.7803	0.1768
history	0.9461	0.7939	0.6825	0.2175
respect	0.9341	0.7969	0.7926	0.1844
news	0.9603	0.6586	0.7021	0.0908
protest	0.8632	0.5885	0.2959	0.1107
discuss	0.7925	0.2079	0.2816	0.0229
party	0.5983	0.1303	0.2440	0.0209

Table 2.3: Class proportions on lower level (LL) and higher level (HL) and class-specific response probabilities for the multilevel measurement model

	$X = 1$ (Maximal)	$X = 2$ (Engaged)	$X = 3$ (Duty)	$X = 4$ (Subject)
$\widetilde{X} = 1$	0.8702	0.1381	0.0235	0.0000
$\widetilde{X} = 2$	0.1156	0.7817	0.1363	0.0059
$\widetilde{X} = 3$	0.0142	0.0794	0.8142	0.1022
$\widetilde{X} = 4$	0.0000	0.0008	0.0260	0.8920

Table 2.4: Estimated misclassification probabilities for step 2 of the three-step approach, with X the true class and $\widetilde{X}$ the assigned class

Table 2.5 presents the estimated covariate effects on lower-level class membership. "Maximal" is the reference class and "high school or less" is the reference level of the covariate. As expected, the three-step slope estimates differ more from the slope estimates for the one-step method and the two-step method compared to the difference between the alternative estimators. Nevertheless, substantively, the results for three approaches are well aligned. When maternal education is higher the probability of belonging to the "Engaged" class relative to the "Maximal" class is smaller, with the greatest effect size in country-level class 3. We can see a similar positive effect, but at a greater magnitude, with respect to the probability of belonging to the "Subject" class relative to the "Maximal" class. It is substantial in country-level class 3. The effect of maternal education on the probability of belonging to the "Duty" class relative to the "Maximal" class is rather weak. The precise point estimates for the three-step estimator are in line with the direction of the effects and the relative effect sizes. We can also see that the difference in the one-step SEs and the two-step SEs is smaller than what we would expect given the simulation results. As such, the pattern of significance of the covariate effects is overall comparable.

In light of the similarity in the substantive results for the methods, we can consider the performance of the three-step methods acceptable even under these suboptimal class separation conditions.

HL Class 1	LL Class 2 (Engaged)		
	one-step	two-step	three-step
matern. educ.	-0.1588***	-0.1536***	-0.1610***
	(0.0417)	(0.0415)	(0.0432)
	LL Class 3 (Duty)		
	one-step	two-step	three-step
matern. educ.	0.1196***	0.1049**	0.0026
	(0.0415)	(0.0438)	(0.0478)
	LL Class 4 (Subject)		
	one-step	two-step	three-step
matern. educ.	-0.5551***	-0.5542***	-0.6117***
	(0.0582)	(0.0592)	(0.0714)
HL Class 2	LL Class 2 (Engaged)		
	one-step	two-step	three-step
matern. educ.	-0.0175	-0.0158	-0.0029
	(0.0434)	(0.0431)	(0.0482)
	LL Class 3 (Duty)		
	one-step	two-step	three-step
matern. educ.	0.1983***	0.1922***	0.1924***
	(0.0463)	(0.0494)	(0.0520)
	LL Class 4 (Subject)		
	one-step	two-step	three-step
matern. educ.	-0.2471***	-0.2517***	-0.3088***
	(0.0732)	(0.0745)	(0.0893)
HL Class 3	LL Class 2 (Engaged)		
	one-step	two-step	three-step
matern. educ.	-0.4458***	-0.4301***	-0.2059***
	(0.0741)	(0.0743)	(0.0782)
	LL Class 3 (Duty)		
	one-step	two-step	three-step
matern. educ.	-0.1262	-0.1316	-0.1444
	(0.1047)	(0.1080)	(0.1269)
	LL Class 4 (Subject)		
	one-step	two-step	three-step
matern. educ.	-1.1527***	-1.1487***	-1.2835***
	(0.1691)	(0.1911)	(0.2537)

Table 2.5: Estimated effect of maternal post-high school education on latent class membership for the three-, two-, and one-step estimators, where class 1 (Maximal) is the reference category, *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1

2.7 Discussion

We have proposed a bias-adjusted three-step ML estimator for multilevel latent class analysis with covariates. In the first step, the single-level LC model without covariates is fitted to the data, ignoring the multilevel structure. In the second step, (lower-level) units are assigned to the classes on the basis of some classification rule, e.g., modal assignment. In the third step, the multilevel LC model with the covariates is estimated while correcting for classification error to obtain unbiased logistic parameter estimates for the association between class membership and covariates.

The performance of the three-step estimator was compared to the currently available modeling options for multilevel LCA with covariates, namely the one-step and two-step estimators, by means of a simulation study. Considering bias, the results showed that the three-step method is a viable alternative in most of the simulation conditions that were considered. Its performance is problematic when the higher-level separation is large and lower-level separation is small, but this bias is reduced substantially when the lower-level separation is at least moderate, and eliminated when the lower-level separation is high. When the higher-level separation is also moderate the three-step method performs as good as identical to the alternative estimators. Considering standard errors, the simulation study showed that the three-step method under-estimates them, which is a known feature of the method in single-level LCA (Bolck et al., 2004; Vermunt, 2010). The bias is less severe when the separation is larger on both the lower level and also larger on the higher level.

The findings for the performance of the three-step bias-correction method in multilevel LCA inform the ongoing debate about the choice between random-effect specifications and fixed-effect specifications. This debate has a long history in latent variable modeling (Aitkin, 1999; Aitkin & Alfó, 1998; Kankaraš, Moors, & Vermunt, 2018) and beyond (for example in econometrics; see e.g. Peracchi, 2001). The reported performance of the random-effect three-step method under different combinations of lower-level separation and higher-level separation can serve as a practical indication of best-practice modeling approaches to applied multilevel LCA users. If the interest is in the lower-level structure, fixed-effect modeling may be a more appealing option when the higher-level separation is stronger relative to the lower-level separation.

In a real data example with moderate class separation at the lower level and between moderate and large class separation at the higher level, we compared the estimated covariate effects and their SEs for the three-step estimator and the alternative one-step and two-step estimators. Specifically, we identified citizenship norms among adolescents and analyzed its association with socioeconomic status. While the estimates were not identical, we observed that the three methods produced well-aligned substantial results, such that the three-step could be considered a legitimate modeling option even under these suboptimal conditions.

The current study is limited in the sense that we analyzed the performance of the proposed three-step method only for situations when the estimated model was correctly specified. One interesting avenue for future research is therefore to investigate the performance of the proposed method when modeling assumptions do not hold. In multilevel contexts, measurement non-invariance on the higher level, in which the item-response probabilities for the lower-level classes differ between higher-level classes, is likely to occur. This may lead to differences in the estimated response probabilities for the multilevel measurement model of interest and the step 1 single-level measurement model, thus introducing additional error in the step 2 class assignment. As such, it is recommended that future research look into controlling for measurement non-invariance on the higher level. For example, measurement non-invariance could be corrected for by means of fitting the multilevel measurement model and deriving the single-level model by marginalizing the conditional lower-level class proportions over the unconditional higher-level class proportions. Alternatively, measurement non-invariance could be corrected for by means of group (i.e., j-identifier) fixed effects (similar to Vermunt & Magidson, 2021a).

Another topic for future research is the correction of the SEs of the covariate effects. In the single-level context, correction methods for the bias-adjusted three-step ML method have been presented by Bakk, Oberski, and Vermunt (2014). The need for correction can be expected to depend on sample size and class separation at the lower level and the higher level. It would be worthwhile investigating how to extend these correction methods to the multilevel case and when the corrections are likely to be necessary in practice.

Finally, we have focused on the fully general parametrization of the multilevel latent class model. In applied research it is sometimes more appealing to adopt more constrained versions, such as random intercepts only and fixed slope parameters. While such models were outside the scope of the current study, it is nevertheless an interesting avenue for future research to examine the performance of the three-step under more constrained parametrizations.

Chapter 3

Two-step multilevel latent class analysis in the presence of measurement non-equivalence

Johan Lyrvall, Zsuzsa Bakk[1]*, Jouni Kuha*[2]*, & Jennifer Oser*[3]

3.1 Introduction

The methodological research question that is considered in this article is the following: How can we estimate multilevel latent class models with covariates when there is non-equivalence of measurement in some of the measurement items, using the two-step method of estimation? How well do these estimates perform? We begin by briefly introducing the key terms in this statement.

Latent class (LC) analysis (Goodman, 1974a; Lazarsfeld & Henry, 1968) is used to classify units into subgroups based on multiple observed categorical variables. The LC model takes these observed variables (*items*) to be indicators of a categorical latent variable of interest (*latent class*). For example, Oser, Hooghe, et al. (2023) used LC analysis to identify types of citizenship norms measured by responses to multiple survey questions about different democratic values.

In applied LC analysis, substantive research questions commonly focus on associations between external predictors, or *covariates*, and the probabilities of belonging to the different latent classes. This is operationalised in terms of regression models for the classes given the covariates. For example, Oser, Hooghe, et al. (2023) used socioeconomic predictors to describe how individuals sort into citizenship norms. The model then combines two elements: a *measurement* model for how the items measure the latent classes, and a *structural model* for how the latent classes depend on the covariates.

Basic LC modelling assumes that the units of analysis are independent of each other. This is insufficient when we have hierarchical data where *lower-level units* (such as individual respondents) are nested (clustered) within *higher-level units* (groups). The nesting can extend to still higher levels, but our discussion is limited to the case of two-level hierarchical data. It is assumed that units in different groups are independent of each other, but that lower-level units within the same group need not be independent even conditional on the covariates.

Within-group dependencies can be accommodated by introducing another latent variable which varies at the higher level. When it is categorical, i.e. a higher-level latent class variable, we have a *multilevel latent class model* (Vermunt, 2003). For example, Di Mari et al. (2023b) used multilevel LC analysis to identify citizenship norms within countries, finding two country-level clusters with different prevalences of the individual-level classes of citizenship norms. The higher-level variable

[1]Department of Methodology and Statistics, Leiden University, Netherlands
[2]Department of Statistics, London School of Economics, United Kingdom
[3]Department of Politics and Government, Ben-Gurion University of the Negev, Israel

is analogous to continuous *random effects* in multilevel models which include such variables (see e.g. Rabe-Hesketh and Skrondal 2022 for examples of them). Multilevel LC models can include covariates as predictors of both higher- and lower-level latent classes. Most often substantive interest is focused on the lower level. For instance, Di Mari et al. (2023b) identified socioeconomic predictors of individual-level norms.

We consider likelihood-based estimation of the models. In standard maximum likelihood (ML) estimation, or *one-step estimation*, all the parameters are estimated simultaneously. In contrast, *stepwise estimation* divides estimation of the measurement model and the structural model into separate steps. The one-step approach has the standard optimality properties of ML estimation, but it also has serious drawbacks (see the discussions in Vermunt 2010 and Bakk and Kuha 2018). Practically, it can be computationally demanding, and will require the same computational effort every time the model is changed and re-fitted. Conceptually, estimating the measurement and structural models together has the disadvantage that they will affect each other. Any changes to the structural model, such as adding or removing covariates or changing their functional form, will also change the estimated measurement model, and hence the implied definition of the latent classes. These changes can be so large that they render comparisons of different structural models effectively meaningless.

Stepwise estimation avoids or reduces the disadvantages of the one-step method. It begins by estimating just the parameters of the measurement model (step 1). Different stepwise methods differ in what happens next. *Three-step estimation* assigns observations to the latent classes based on the estimated measurement model (step 2), and then fits the structural model for these assigned classes (step 3). *Bias-adjusted three-step estimation* employs further adjustments to correct for misclassication bias that would arise from naive use of step 2 (see the review in Bakk and Kuha 2021 and references therein).

In contrast, stepwise *two-step estimation* does not assign predicted latent classes, but estimates (in its step 2) the structural model directly from a likelihood where the measurement-model parameters are fixed at their estimates from step 1. Two-step estimation for LC models was first proposed by Bandeen-Roche, Miglioretti, Zeger, and Rathouz (1997) and Xue and Bandeen-Roche (2002), and further developed by Bakk and Kuha (2018). The same idea can also be applied to latent variable models which have continuous rather than categorical latent variables (Rosseel and Loh 2024; Kuha and Bakk 2023).

For multilevel LC models, stepwise methods have been proposed using a bias-adjusted three-step (Lyrvall et al., 2024), an intermediate "two-stage" (Bakk et al., 2022), and the two-step approaches (Di Mari et al., 2023b). We regard the two-step method as the preferred approach because of its simplicity and good performance in previous studies.

A latent variable model has the property of *measurement equivalence* if the measurement model for the items depends *only* on the latent variables but not on any covariates or observed response variables. Violation of this, where measurement is affected also by observed external variables, is known as *measurement non-equivalence*, also known as non-invariance of measurement or differential item functioning (DIF). It can arise, for example, in cross-national surveys from differences in translation or in educational testing from differences in familiarity of test questions for different groups of students which are unrelated to their ability. In the illustrative example that we consider in Section 3.5 of this paper, we allow for possible non-equivalence in survey questions on citizenship norms which may arise from differences in the salience of different civic activities in countries with higher or lower levels of political freedom. There is a large literature on issues of non-equivalence in different applications and for different types of latent variable models (see e.g. Millsap 2011 and Kankaraš, Vermunt, and Moors 2011, and references therein). Masyn (2017) discusses it for LC models, and provides definitions and model specifications.

If there is non-equivalence in the measurement, estimation which ignores this will yield biased estimates also for the structual model. Studies by Asparouhov and Muthén (2014), Janssen, Van Laar, De Rooij, Kuha, and Bakk (2019) and Di Mari and Bakk (2018) show that this bias can be large for latent class models. It is thus often crucial to correctly account for any non-equivalence in model specification and estimation.

One-step estimation in this situation is still standard ML estimation, now for a model which includes covariates also in the measurement model. For stepwise methods, Vermunt and Magidson (2021a) described how bias-adjusted three-step estimation can be implemented for single-level LC models with non-equivalence of measurement. Their key point is to specify the model for its step 1 correctly. This should include those covariates which affect the measurement model, and include them in both the measurement model and the structural model (they should then also be appropriately accounted for in steps 2 and 3).

Vermunt and Magidson (2021a) also note that what they propose for three-step estimation would also be the correct form for step 1 of the two-step method. In this paper we follow up on that point. We combine the elements from previous literature described above, and extend them to develop two-step estimation which allows for non-equivalence of measurement and which can be applied to single-level and multilevel LC models.

The model is defined in Section 3.2 of the paper, and in Section 3.3 we describe how the estimation is implemented. We then evaluate the performance of the method through simulation studies in 3.4 and illustrate it further with an empirical example in Section 3.5.

3.2 Multilevel latent class model with covariates and measurement non-equivalence

Here we give a formal definition of the model that was outlined in Section 3.1. We define its elements in steps, finishing with the introduction of non-equivalence to the measurement model.

Consider hierarchical data where *lower-level units* (individuals) $j = 1, \ldots, n_i$ are nested in *higher-level units* (groups) $i = 1, \ldots, I$. Let Y_{ijh}, $h = 1, \ldots, H$, be the values of H observed variables (*items*) for lower-level unit j in higher-level unit i, and define $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijH})'$. Here each Y_{ijh} is a categorical variable, with possible values $r = 1, \ldots, R_h$. Let $\mathbf{Z}_{ij} = (\mathbf{Z}_i^{H\prime}, \mathbf{Z}_{ij}^{L\prime})'$ be a vector of observed covariates, where the variables in $\mathbf{Z}_{ij}^L$ (*lower-level covariates*) can vary between different lower-level units within the same higher-level unit but $\mathbf{Z}_i^H$ (*higher-level covariates*) vary only between the higher-level units. We take $\mathbf{Z}_i^H$ to include a constant 1, thus introducing an intercept term to all the regression models described below.

The items $\mathbf{Y}_{ij}$ are regarded as observed indicators of a discrete latent variable X_{ij} with categories (*latent classes*) $t = 1, \ldots, T$. The standard latent class (LC) model specifies the joint probability function of X_{ij} and $\mathbf{Y}_{ij}$ as $P(\mathbf{Y}_{ij}, X_{ij}) = P(X_{ij})\, P(\mathbf{Y}_{ij}|X_{ij})$. This has two basic elements, the *structural model* $P(X_{ij})$ for the probabilities of the latent classes, and the *measurement model* $P(\mathbf{Y}_{ij}|X_{ij})$ for how the items measure the latent classes. We make throughout the assumption, which is standard in LC analysis, that Y_{ijh} for different h are conditionally independent of each other given the latent class. The measurement model can then be written as

$$P(\mathbf{Y}_{ij}|X_{ij}) = \prod_{h=1}^{H} P(Y_{ijh}|X_{ij}). \tag{3.1}$$

Next, the model is extended to accommodate the hierarchical structure of the data. This is done by expanding the structural model to $P(X_{ij}, W_i) = P(W_i)\, P(X_{ij}|W_i)$, where W_i is another categorical latent class variable, with categories $m = 1, \ldots, M$. It varies only between higher-level units i, so we refer to it as the *higher-level LC variable* and X_{ij} as the *lower-level LC variable.* It is assumed that $\mathbf{Y}_{ij}$ and W_i are conditionally independent given X_{ij}, and that X_{ij} for the same i are conditionally independent given W_i. Averaged over $P(W_i)$, however, values of X_{ij} for different j within the same group i will be associated because they share the same W_i. In this sense, W_i is a categorical analogy of continuous random effects in multilevel (random effects) models, and the model is referred to as a *multilevel* (here two-level) *LC model.*

We then introduce covariates to the structural model, as

$$P(X_{ij}, W_i|\mathbf{Z}_{ij}) = P(W_i|\mathbf{Z}_i^H)\, P(X_{ij}|W_i, \mathbf{Z}_{ij}), \tag{3.2}$$

noting that higher-level classes W_i can only depend on higher-level covariates $\mathbf{Z}_i^H$ but lower-level classes X_{ij} can depend on both lower- and higher-level covariates. We specify these models as the multinomial logistic models

$$P(W_i = m|\mathbf{Z}_i^H) = \frac{\exp(\boldsymbol{\alpha}_m' \mathbf{Z}_i^H)}{\sum_{l=1}^{M} \exp(\boldsymbol{\alpha}_l' \mathbf{Z}_i^H)} \quad \text{and} \tag{3.3}$$

$$P(X_{ij} = t|W_i = m, \mathbf{Z}_{ij}) = \frac{\exp(\boldsymbol{\gamma}_{t|m}' \mathbf{Z}_{ij})}{\sum_{s=1}^{T} \exp(\boldsymbol{\gamma}_{s|m}' \mathbf{Z}_{ij})}, \tag{3.4}$$

where $\boldsymbol{\alpha}_m$ and $\boldsymbol{\gamma}_{t|m}$ for $m = 1, \ldots, M$ and $t = 1, \ldots, T$ are parameter vectors, and $\boldsymbol{\alpha}_1 = \mathbf{0}$ and $\boldsymbol{\gamma}_{1|m} = \mathbf{0}$ for all m for identifiability. The specification may include constraints on the parameters, for example when some of them are 0 or when matching elements of $\boldsymbol{\gamma}_{t|m}$ are equal for all m. Often the focus of substantive interest is on model (3.4) for the lower-level latent class X_{ij}, and the higher-level class W_i is regarded just as a random effect to allow for within-group associations between X_{ij}. In that case, model (3.3) will often include just the intercept terms $\boldsymbol{\alpha}_m = \alpha_m$.

The model defined by (3.1) and (3.2) is a standard multilevel LC model with covariates (Vermunt 2003;Bakk et al. 2022; Di Mari et al. 2023b;Lyrvall et al. 2024). A key feature of it is that the measurement model (3.1) does not depend on $\mathbf{Z}_{ij}$. This can be relaxed by introducing covariates also to this, as

$$P(\mathbf{Y}_{ij}|X_{ij}, \mathbf{Z}_{ij}) = \prod_{h=1}^{H} P(Y_{ijh}|X_{ij}, \mathbf{Z}_{ijh}^*),$$

where the models for the individual items are multinomial logistic models

$$P(Y_{ijh} = r|X_{ij} = t, \mathbf{Z}_{ijh}^*) = \frac{\exp(\boldsymbol{\delta}_{hr|t}' \mathbf{Z}_{ijh}^*)}{\sum_{q=1}^{R_h} \exp(\boldsymbol{\delta}_{hq|t}' \mathbf{Z}_{ijh}^*)} \tag{3.5}$$

for $r = 1, \ldots, R_h$, and $\boldsymbol{\delta}_{hq|t}$ are parameter vectors with $\boldsymbol{\delta}_{h1|t} = \mathbf{0}$ for all h, t. This kind of measurement model for item Y_h is *non-equivalent* with respect to the covariates in $\mathbf{Z}_{ijh}^*$. We write this with the subscript h to denote only those elements of $\mathbf{Z}$ which do affect the measurement model for the hth item. This is useful for clarity, because it is very common that these include only a subset of the variables in $\mathbf{Z}$, and that they are different for different items. There may be parameter constraints, for example so that the coefficients of $\mathbf{Z}_{ijh}^*$ (except for the intercept) do not depend on latent class t, or that even for the same h they may be non-zero for some latent classes but zero for others. If $\mathbf{Z}_{ijh}^*$ includes only the constant 1, measurement of item Y_{ijh} is *equivalent* with respect to all of the covariates.

Let $\mathbf{Y}_i = (\mathbf{Y}_{i1}', \ldots, \mathbf{Y}_{in_i}')'$ and $\mathbf{Z}_i = (\mathbf{Z}_{i1}', \ldots, \mathbf{Z}_{in_i}')'$ denote all the observed values of the items and the covariates for higher-level unit i. The model for these observed data is obtained by averaging over the distributions of the latent W_i and X_{ij}, as

$$P(\mathbf{Y}_i|\mathbf{Z}_i; \boldsymbol{\theta}) = \sum_{m=1}^{M} P(W_j = m|\mathbf{Z}_i^H; \boldsymbol{\theta}_2) \prod_{j=1}^{n_i} \left\{ \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}; \boldsymbol{\theta}_2) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t, \mathbf{Z}_{ijh}^*; \boldsymbol{\theta}_1) \right\} \tag{3.6}$$

where we have also introduced parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ into the notation. Here $\boldsymbol{\theta}_1$ denotes all the parameters of the measurement model, i.e. the δs in (3.5), and $\boldsymbol{\theta}_1$ all the parameters of the structural model, i.e. the αs and γs in (3.3) and (3.4).

Model (3.6) is a multilevel (here two-level) latent class model with covariates and with non-equivalence of measurement. What we examine in this paper is two-step methods of estimating the parameters of this model, with focus on the structural parameters $\boldsymbol{\theta}_2$. In the general presentation of the method in Section 3.3 we take the choice of $\mathbf{Z}^*_{ij1}, \ldots, \mathbf{Z}^*_{ijH}$ as given, i.e. we assume that it has already been determined which covariates are needed to allow for non-equivalence of measurement in different items. Model selection procedures for deciding on this are described by Masyn (2017) and Vermunt and Magidson (2021a); an illustration of them is included in our applied example in Section 3.5. We also assume that the specification of the measurement model is such that the parameters of the structural model are formally and practically identified. This requires, in essence, that the non-equivalence should not be too extensive, at a minimum that it does not affect all of the items in $\mathbf{Y}_{ij}$.

3.3 Two-step estimation of the model parameters

The $\mathbf{Y}_i$ for different higher-level units i are taken to be conditionally independent given $\mathbf{Z}_i$. The log-likelihood function for the model that was defined in Section 3.2 can then be written as $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{I} \log P(\mathbf{Y}_i|\mathbf{Z}_i; \boldsymbol{\theta})$, where $P(\mathbf{Y}_i|\mathbf{Z}_i; \boldsymbol{\theta})$ is given by (3.2) combined with (3.3)–(3.5).

One-step maximum likelihood (ML) estimates of the parameters are obtained by maximizing $\ell(\boldsymbol{\theta})$ with respect to all of $\boldsymbol{\theta}$ at once. In contrast, two-step estimation divides the estimation into two steps. In its step 1, an estimate $\tilde{\boldsymbol{\theta}}_1$ of the measurement parameters is obtained. In step 2, estimates $\tilde{\boldsymbol{\theta}}_2$ of the structural parameters are obtained by maximizing $\ell(\tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$, i.e. using the same log-likelihood as for one-step estimation but treating now the measurement parameters $\boldsymbol{\theta}_1$ fixed at their estimated values $\tilde{\boldsymbol{\theta}}_1$ from step 1.

This idea of two-step estimation has been examined for single-level latent class models by Bakk and Kuha (2018) and for multilevel LC models by Di Mari et al. (2023b). What is new here is that we want to extend it to the case where the model includes non-equivalence of measurement. The key question is then how step 1 should be carried out. The general answer is that it should use the simplest model that allows valid estimation of $\boldsymbol{\theta}_1$. To present this, we write now $\mathbf{Z}_{ij} = (\mathbf{Z}^{\dagger\prime}_{ij}, \mathbf{Z}^{*\prime}_{ij})'$, where $\mathbf{Z}^*_{ij}$ denotes the union of $\mathbf{Z}^*_{ijh}$ over h, i.e. those covariates that appear in the measurement model for at least one item, and $\mathbf{Z}^\dagger_{ij}$ denotes those covariates that do not appear anywhere in the measurement model. Let $p(\mathbf{Z}^\dagger_{ij}|\mathbf{Z}^*_{ij})$ denote the conditional joint distribution of $\mathbf{Z}^\dagger_{ij}$ given $\mathbf{Z}^*_{ij}$. The conditional distribution for the latent class variables and the items given $\mathbf{Z}^*_{ij}$ only is obtained by marginalising over this, as

$$
\begin{aligned}
&P(\mathbf{Y}_{ij}, X_{ij}, W_i|\mathbf{Z}^*_{ij}; \boldsymbol{\theta}_1, \boldsymbol{\theta}^*_2) \\
&= \left[\int P(X_{ij}, W_i|\mathbf{Z}^\dagger_{ij}, \mathbf{Z}^*_{ij}; \boldsymbol{\theta}_2)\, p(\mathbf{Z}^\dagger_{ij}|\mathbf{Z}^*_{ij})\, d\mathbf{Z}^\dagger_{ij} \right] P(\mathbf{Y}_{ij}|X_{ij}, \mathbf{Z}^*_{ij}; \boldsymbol{\theta}_1) \\
&= P(X_{ij}, W_i|\mathbf{Z}^*_{ij}; \boldsymbol{\theta}^*_2)\, P(\mathbf{Y}_{ij}|X_{ij}, \mathbf{Z}^*_{ij}; \boldsymbol{\theta}_1) \\
&= P(W_i|\mathbf{Z}^{H*}_i; \boldsymbol{\theta}^*_2)\, P(X_{ij}|W_i, \mathbf{Z}^*_{ij}; \boldsymbol{\theta}^*_2)\, P(\mathbf{Y}_{ij}|X_{ij}, \mathbf{Z}^*_{ij}; \boldsymbol{\theta}_1).
\end{aligned}
\tag{3.7}
$$

This is of the same multilevel LC form as the full model given $\mathbf{Z}_{ij}$ which led to (3.6). The two have different structural models, since (3.7) is conditional on $\mathbf{Z}^*_{ij}$ only (so we denote its structural parameters by $\boldsymbol{\theta}^*_2$ rather than $\boldsymbol{\theta}_2$). Crucially, however, both have the same measurement model $P(\mathbf{Y}_{ij}|X_{ij}, \mathbf{Z}^*_{ij}; \boldsymbol{\theta}_1)$, with the same $\boldsymbol{\theta}_1$. The measurement parameters $\boldsymbol{\theta}_1$ can thus be estimated from this, using an observed-data log likelihood that is obtained by marginalising (3.7) over X_{ij} and W_i. This is the key result that was derived by Vermunt and Magidson (2021a) for step 1 of three-step estimation for single-level LC models, and it holds also for two-step estimation for the multilevel models that we consider here. Vermunt and Magidson (2021a) also observed that the same result holds even if the model includes observed variables that are treated as distal outcomes rather than covariates, even when they depend on the items $\mathbf{Y}_{ij}$; this is because they would be integrated out from an expression like (3.7). If the model has full measurement equivalence, i.e. $\mathbf{Z}^*_{ij}$ includes only the constant 1, (3.7) integrates out all the covariates. The step-1 model is then a multilevel LC model without covariates, as in Di Mari et al. (2023b).

We note that this derivation involves one approximation. This is that if the structural models given $\mathbf{Z}_{ij}$ are multinomial logistic models as in (3.3) and (3.4), they will in general be only approximately of a multinomial logistic form given a smaller set $\mathbf{Z}^*_{ij}$ (unless this is empty or includes only a single categorical variable). We do not expect that this will have a meaningful impact on the quality of the step-1 estimates of $\boldsymbol{\theta}_1$ (we note also that the same inconsistency arises whenever any multinomial logistic models are fitted given different sets of covariates, even for observed response variables).

In summary, when there is non-equivalence of measurement with respect to covariates $\mathbf{Z}^*_{ij}$, step 1 of two-step estimation should be for a model which includes these $\mathbf{Z}^*_{ij}$ in both the structural model and the measurement model. This is still simpler than one-step estimation if $\mathbf{Z}^*_{ij}$ is smaller than the full set of covariates $\mathbf{Z}_{ij}$. Estimates $\tilde{\boldsymbol{\theta}}_1$ of the measurement parameters from this step 1 are carried forward to step 2 (and estimates of the structural parameters $\boldsymbol{\theta}^*_2$ are discarded). Two-step estimates $\tilde{\boldsymbol{\theta}}_2$ of the structural parameters are then obtained from step 2 by maximizing $\ell(\tilde{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$ with respect to $\boldsymbol{\theta}_2$.

For estimation of standard errors of $\tilde{\boldsymbol{\theta}}_2$, two broad approaches are possible. One of them accounts for sampling uncertainty in $\tilde{\boldsymbol{\theta}}_1$ by including a term corresponding to this in the standard error calculation (Bakk and Kuha 2018; Di Mari et al. 2023b). The other, simpler approach, omits this term, in effect taking the estimated measurement model from step 1 as an a priori fixed definition of the latent classes (see Kuha and Bakk 2023 for a discussion of these options). In our applied example in Section 3.5 we use this simpler approach to calculate the standard errors.

3.4 Simulation study

3.4.1 Design

We use a simulation study to examine the performance of the proposed two-step method of estimation for multilevel latent class models with measurement non-equivalence (abbreviated *MNE* below). We focus on results for estimated parameters of the structural model for the lower-level classes (model (3.4) in Section 3.2), because this is typically the focus of substantive research questions in applications of multilevel LC models. Our primary question of interest is how well the estimates perform when MNE is correctly specified in the measurement model, and a secondary question is how much bias they have when MNE is incorrectly ignored and equivalence of measurement is assumed. For both of these questions, we also use one-step estimation as a comparator.

Two main factors are varied in the simulation settings: separation of the latent classes (i.e. the strength of the measurement model) and magnitude of the MNE. It is well known for models without MNE that estimates behave better when the classes are more clearly separated (Bakk & Kuha, 2018; Di Mari et al., 2023b; Lyrvall et al., 2024; Vermunt, 2010), and we would expect the same to be the case here. Similarly, we expect that estimation is more demanding if non-equivalence is more pronounced. A question of interest is then how large these differences may be.

Each simulated sample has $I = 100$ higher-level units i and $n_i = 100$ lower-level units j in each i. Each higher-level unit belongs to one of two known groups, identified by an observed variable $G_i = 0, 1$. The value of G_i is drawn at random for each i, with probability $P(G_i = 1) = 0.5$. Non-equivalence of measurement may exist between these groups. This structure might correspond, for example, to a multicultural educational study where the higher-level units are schools, lower-level units are students, and the two groups are two different languages of instruction in the schools.

We consider models with $T = 3$ lower-level latent classes (categories of X_{ij}) and $M = 2$ higher-level latent classes (categories of W_i). Model (3.3) for W_i has no covariates, i.e. $\mathbf{Z}^H_i = 1$, and we set $P(W_i = 1) = 0.6$ and $P(W_i = 2) = 0.4$. Model (3.4) for X_{ij} has G_i as its only covariate, i.e. i.e. $\mathbf{Z}_{ij} = (1, G_i)'$. The intercepts of this model are set so that, averaged over the distribution of G_i, we have $P(X_{ij} = 1|W_i = 1) = P(X_{ij} = 3|W_i = 2) = 0.18$, $P(X_{ij} = 2|W_i = 1) = P(X_{ij} = 2|W_i = 2) = 0.31$, and $P(X_{ij} = 3|W_i = 1) = P(X_{ij} = 1|W_i = 2) = 0.51$.

In all of the simulations, in the model for X_{ij} all coefficients of G_i (i.e. in all $\boldsymbol{\gamma}_{t|m}$ in (3.4) for $t = 2, 3$ and $m = 1, 2$) are equal to 0.5. The estimated model correctly assumes that these coefficients do not vary by the higher-level class m, so that the model has two estimable coefficients of G_i. These are the parameters we focus on, considering all of their estimates together.

The lower-level latent class is measured by $H = 6$ binary items Y_{ijh} for $h = 1, \ldots, H$, each with values 0 and 1. Consider the item response probabilities $\pi_{h(t)g} = P(Y_{ijh} = 1|X_{ij} = t, G_i = g)$. Here for simplicity we write G_i in place of the covariates $\mathbf{Z}^*_{ijh}$ because in all cases where there is MNE we have $\mathbf{Z}^*_{ijh} = (1, G_i)$ (and when there is no MNE, $\mathbf{Z}^*_{ijh} = 1$ and $\pi_{h(t)0} = \pi_{h(t)1}$). In all settings $\pi_{h(t)g}$ has a high value (> 0.5) for all items h =1–6 in the first lower-level class ($t = 1$), for items 1–3 in class $t = 2$ and for no items in class $t = 3$, and low probabilities (≤ 0.5) otherwise. In different simulations we then allow MNE by group G_i for some of the $\pi_{h(t)g}$. The strength of class separation and magnitude of MNE are determined by how these probabilities vary and how far they are from 0.5.

We consider simulation conditions with weaker and stronger lower-level class separation separately for low and high values of $\pi_{h(t)g}$, resulting in four settings for class separation. These are combined with three conditions for MNE — none, weak and strong — resulting in 12 simulation conditions in total. When there is MNE, it affects the measurement models of some items in latent classes 1 and 2 but none of them in class 3. In the weaker MNE condition, classes 1 and 2 have MNE for items $h = 1, 2$. In the stronger condition, class 1 has MNE in items 1–4 and class 3 in items 1–3. Thus MNE here affects only those probabilities $\pi_{h(t)g}$ that are greater than 0.5. In each case its effect is to shift the response probability down by 0.1 for group 1, i.e. $\pi_{h(t)1} = \pi_{h(t)0} - 0.1$. The resulting values of the response probabilities in the twelve simulation conditions are summarised in Table 3.1.

Patterns of response probabilities:

		Response probability $\pi_{h(t)g} = P(Y_h = 1|X = t, G = g)$ for item (h)					
Class (t)	Group (g)	1	2	3	4	5	6
1	0	H_{0a}	H_{0a}	H_{0b}	H_{0b}	H	H
	1	H_{1a}	H_{1a}	H_{1b}	H_{1b}	H	H
2	0	H_{0a}	H_{0a}	H_{0b}	L	L	L
	1	H_{1a}	H_{1a}	H_{1b}	L	L	L
3	0	L	L	L	L	L	L
	1	L	L	L	L	L	L

Values of the probabilities in different simulation conditions:

Condition	Separation (low $\pi_{h(t)g}$)	Separation (high $\pi_{h(t)g}$)	Measurement non-equiv.	(H_{0a}, H_{1a})	(H_{0b}, H_{1b})	H	L
1	Weak	Weak	None	0.8	0.8	0.8	0.5
2	Weak	Strong	None	0.9	0.9	0.9	0.5
3	Strong	Weak	None	0.8	0.8	0.8	0.2
4	Strong	Strong	None	0.9	0.9	0.9	0.1
5	Weak	Weak	Weak	(0.8, 0.7)	0.8	0.8	0.5
6	Weak	Strong	Weak	(0.9, 0.8)	0.9	0.9	0.5
7	Strong	Weak	Weak	(0.8, 0.7)	0.8	0.8	0.2
8	Strong	Strong	Weak	(0.9, 0.8)	0.9	0.9	0.1
9	Weak	Weak	Strong	(0.8, 0.7)	(0.8, 0.7)	0.8	0.5
10	Weak	Strong	Strong	(0.9, 0.8)	(0.9, 0.8)	0.9	0.5
11	Strong	Weak	Strong	(0.8, 0.7)	(0.8, 0.7)	0.8	0.2
12	Strong	Strong	Strong	(0.9, 0.8)	(0.9, 0.8)	0.9	0.1

Table 3.1: Values of the item response probabilities in different conditions considered in the simulations. In the lower table, two values for (H_{0a}, H_{1a}) and/or (H_{0b}, H_{1b}) indicate that the values of these probabilities are different in groups $g = 0, 1$, i.e. that there is measurement non-equivalence in the corresponding part of the model.

For each of the conditions, we generate 250 random samples. The data analysis is carried out in `Mplus` (Muthén & Muthén, 2017) and `R` (R Core Team, 2024), using the package `MplusAutomation` (Hallquist & Wiley, 2018).

3.4.2 Results

Tables 3.2 and 3.3 show the simulation results, in the form of the average bias, root mean squared error (RMSE) and median absolute error (MAE) of estimates over the 250 simulations in each of the simulation scenarios. As noted above, the parameters considered here are the two coefficients of G_i in the model for the lower-level class X_{ij}, both with the true value of 0.5. We consider their estimates together, so that we have 500 estimated values for each simulation setting.

Class separation for		True level of measurement non-equivalence					
		None		Weak		Strong	
(low $\pi_{h(t)g}$)	(high $\pi_{h(t)g}$)	One-step	Two-step	One-step	Two-step	One-step	Two-step
Mean bias:							
Weak	Weak	0.001	−0.012	0.263	0.055	0.743	0.256
Weak	Strong	0.003	−0.002	0.031	0.073	0.153	0.258
Strong	Weak	−0.003	−0.003	0.028	0.027	0.128	0.126
Strong	Strong	0.000	−0.001	0.009	0.008	0.058	0.057
Root mean squared error:							
Weak	Weak	0.125	0.121	1.238	0.412	1.888	0.741
Weak	Strong	0.088	0.086	0.312	0.267	0.520	0.480
Strong	Weak	0.067	0.067	0.137	0.135	0.225	0.218
Strong	Strong	0.058	0.058	0.079	0.079	0.118	0.116
Median absolute error:							
Weak	Weak	0.084	0.080	0.533	0.361	0.930	0.683
Weak	Strong	0.055	0.057	0.288	0.244	0.514	0.438
Strong	Weak	0.048	0.048	0.116	0.114	0.181	0.168
Strong	Strong	0.036	0.036	0.059	0.059	0.090	0.089

Table 3.2: Estimation assuming full equivalence of measurement. Mean bias, root mean squared error (RMSE) and median absolute error (MAE) of two-step and one-step estimates of the structural parameters. The results are across the 2×250 estimates of two coefficients of the covariate G (both with true value of 0.5) in the model for lower-level latent class X, over 250 simulation replications in each of the twelve simulation conditions in Table 3.1.

Consider first estimation where measurement non-equivalence is ignored, i.e. when both two-step and one-step estimates are calculated under the assumption of full equivalence of measurement. These results are shown in Table 3.2. When the true model has no MNE, there is little difference between the two estimators and both are essentially unbiased. Both of them become increasingly seriously biased when the true measurement model involves increasing levels of MNE. This bias is also larger when class separation is weaker, i.e. when the measurement model is weak. Here there are also noticeable differences between the two estimators, in that the two-step estimates have mostly smaller bias and smaller RMSE than the one-step estimates, especially in the more difficult low-separation settings. The same is true for MAEs, showing that the poorer performance of the one-step estimates is fairly general and not just due to a small number of extreme values of the estimates.

Table 3.3 shows the results in the eight simulation conditions where MNE is present, when the estimators are based on a correct specification for the MNE. Both estimators again perform better when the separation between the latent classes is stronger. This is as expected, and consistent with previous results for estimation in situations with no MNE (e.g. Vermunt 2010; Bakk and Kuha 2018). Here the most challenging conditions are the ones where low item response probabilities (i.e. the ones indicated by 'L' in Table 3.1) are 0.5, so that they are not very clearly distinguished from the higher response probabilities. The estimators perform reasonably well, and in most cases essentially similarly. However, some differences between them emerge when class separation is weak and there

Class separation for		Level of measurement non-equivalence			
		Weak		Strong	
(low $\pi_{h(t)g}$)	(high $\pi_{h(t)g}$)	One-step	Two-step	One-step	Two-step
Mean bias:					
Weak	Weak	0.002	−0.021	−0.066	−0.118
Weak	Strong	0.003	−0.011	−0.008	−0.101
Strong	Weak	−0.003	−0.006	−0.002	−0.056
Strong	Strong	−0.001	−0.001	0.004	−0.007
Root mean squared error:					
Weak	Weak	0.159	0.127	0.560	0.187
Weak	Strong	0.098	0.086	0.222	0.134
Strong	Weak	0.072	0.066	0.103	0.087
Strong	Strong	0.059	0.058	0.060	0.054
Median absolute error:					
Weak	Weak	0.101	0.086	0.287	0.124
Weak	Strong	0.069	0.059	0.139	0.109
Strong	Weak	0.049	0.045	0.067	0.064
Strong	Strong	0.037	0.035	0.041	0.035

Table 3.3: Estimation under correctly specified model for measurement non-equivalence (MNE). Mean bias, root mean squared error (RMSE) and median absolute error (MAE) of two-step and one-step estimates of the structural parameters. The results are across the 2×250 estimates of two coefficients of the covariate G (both with true value of 0.5) in the structural model for lower-level latent class X, over 250 simulation replications in each of the eight simulation conditions in Table 3.1 that involve MNE.

is a large amount of MNE. Here the two-step estimates have a little more bias, but clearly lower RMSE and MAE than the one-step estimates. In these most difficult situations a large proportion of the one-step estimates are thus quite far from the true parameters, whereas two-step estimation substantially reduce these extremes.

3.5 Empirical example

We illustrate the proposed two-step method of estimation for multilevel LC models with non-equivalence of measurement with an analysis of cross-national data on citizenship norms among adolescents. The data come from the International Civic and Citizenship Education Study 2016 (Schulz et al., 2018), which was conducted by the International Association for the Evaluation of Educational Achievement, and are accessed via the `R` package `multilevLCA` (Lyrvall et al., 2023). These data have been used in previous substantive studies of citizenship norms (Hooghe & Oser, 2015; Hooghe et al., 2016; Oser & Hooghe, 2013; Oser, Hooghe, et al., 2023). For details on data cleaning and recoding, see [reference with DOI to be added].

The survey asked 14-year-old adolescents to state their level of agreement on whether a set of activities are important for a person to be considered a good adult citizen. We include responses to five such questions, related to activities that correspond to engaged citizenship: participation in local activities (we label this item *local*), engagement in political conversations (*discuss*), show of support for environmental protection activities (*envir*), promotion of human rights (*rights*), and participation in peaceful protests (*protest*). The responses are coded in a binary form, as 1 if the respondent regarded the activity as very or quite important for being a good adult citizen, and 0 if they thought it not very or not at all important.

These five binary variables are the measurement items (Y_{ijh} in the notation above). Individual-level latent classes (X_{ij}) measured by them will characterise different profiles of what an adolescent considers important in a good citizen. We have hierarchical data where individual children (lower-level units j) are nested within countries (higher-level units i). We consider structural models where the proportions of X_{ij} may vary by two country characteristics (non-constant covariates in $\mathbf{Z}_{ij} = \mathbf{Z}_i$), the country's wealth and its civic freedom, specifically press freedom. We do not include covariates for the higher-level latent classes W_i, so $\mathbf{Z}_i^H$ in the notation of Section 3.2 includes only a constant. Wealth is measured by logarithm of gross domestic product in U.S. dollars (covariate *lnGDPusd*), and a covariate on press freedom is based on the 2016 World Press Freedom Index (PFI) by Reporters Without Borders. Civic freedom has previously not been considered as an explanatory variable in the latent class analysis citizenship norms literature. For clarity of this illustrative example, we consider data from two groups of countries which have very different levels of press freedom. Five of the countries are among those with the highest levels of PFI — Finland (ranked 1st), Netherlands (2), Norway (3), Denmark (4), and Sweden (8) — and three among the lowest — Colombia (134), Russia (148), and Mexico (149). We define a dummy variable (*lowPFI*) which is 1 for the countries in the low-PFI group and 0 for the high-PFI group[4]. The sample sizes range from 2,728 (Netherlands) to 7,138 (Russia), with a total combined sample of 40,837 respondents.

We also allow for the possibility of MNE in some of the items, with respect to *lowPFI*. The two groups of countries defined by it have very different constraints on political expression, and the different activities mentioned in the survey items may have different relative salience for adolescents' perceptions on what it takes to be a good citizen. In particular, we speculate that this may be the case for support for environmental protection, promotion of human rights, and participation in peaceful protests, which are more public and/or politically contentious activities. We therefore consider the possibility of MNE in these items. The citizenship norms literature has not previously analyzed civic freedom as a potential confounding variable in the identification of latent classes.

We first identified the optimal number of latent classes. This was based on the Bayesian information criterion (BIC) combined with considerations of substantive clarity of the estimated LC structure. A general recommendation is to perform this first step of model selection without covariates and under the assumption of equivalence of measurement (Masyn, 2017). We first estimated single-level models with one to five latent classes, and concluded that the four-class specification was preferred. We then estimated two-level LC models, with individual countries as the higher-level units, still with equivalence of measurement and without covariates. With four lower-level classes, the best BIC value was obtained for a model with three higher-level classes. This multilevel model is preferred to the four-class single-level model, indicating that allowing for the hierarchical clustering structure is desirable. We select the two-level model with four high-level and three low-level classes for the rest of the modeling.

In the second step of model selection, we add MNE with respect to *lowPFI* to this multilevel model. We consider it for all combinations of the three items *envir*, *rights*, and *protest*, both when allowing MNE to vary across classes and when restricting MNE to be invariant (on the logit scale) across classes (i.e. constraining the coefficient of *lowPFI* in $\boldsymbol{\delta}_{hr|t} = \boldsymbol{\delta}_{hr}$ in (3.5) not to depend on latent class t). Here *lowPFI*, i.e. the covariate in $\mathbf{Z}_{ij}^*$, is included also in the model for the latent class variable X_{ij}. The best BIC value is obtained for a model which includes class-invariant MNE in two items, *envir* and *protest*. In particular, it is preferred to a model with full equivalence of measurement. This indicates that MNE is present in the data.

Estimates of the measurement model parameters $\boldsymbol{\theta}_1$ for the selected model from this step are also the step-1 estimates of these parameters for two-step estimation, as discussed in Section 3.3. The item response probabilities implied by this model are shown in Table 3.4. The first class places importance on all five items. The second class emphasizes the items related to specific topics (*local*, *envir*, *rights*), but not as much or at all the ones related to method of engagement (*discuss*, *protest*). Individuals belonging to the third class have middling probabilities of endorsing each of the items, and those in the fourth class do not place importance on any of them as criteria for a good adult citizen. We label class 1 *Maximal*, class 2 *Topic*, class 3 *Medium*, and class 4 *Unengaged*. The same interpretation of the classes would also be obtained from a model which constrains the measurement models to be fully equivalent, item probabilities from which are also shown in Table 3.4

[4]An alternative analytical approach would be to use the original continuous PFI score, which is ranging from 0 to 100. In this empirical example, we focus on the binary *low-high* classification for ease of interpretation. Because the variation in PFI score between the five countries is substantially larger between these two groups than within these groups, we expect this choice of analytical approach has little qualitative impact on the results (among the low-PFI countries, the PFI scores are 55.89, 50.97, and 50.67 for Colombia, Russia, and Mexico, respectively; among the high-PFI countries, they are 91.41, 91.24, 91.21, 91.11, and 87.67 for Finland, Netherlands, Norway, Denmark, and Sweden, respectively).

for comparison. The implications for allowing for MNE are seen in the probabilities for items *envir* and *protest* in the selected model. Here in all classes the probabilities of endorsing these items are higher in countries with low press freedom. In other words, adolescents in countries with low levels of press freedom are more uniformly likely to regard support for environmental protection and participation in peaceful protests as characteristics of a good adult citizen than are adolescents in countries with more press freedom.

	Selected model				Model with full measurement equivalence			
Item	Cl. 1 *Maximal*	Cl. 2 *Topic*	Cl. 3 *Medium*	Cl. 4 *Uneng.*	Cl. 1 *Maximal*	Cl. 2 *Topic*	Cl. 3 *Medium*	Cl. 4 *Uneng.*
local	0.981	0.961	0.600	0.096	0.980	0.961	0.660	0.113
discuss	0.953	0.000	0.294	0.074	0.981	0.001	0.298	0.091
rights	0.988	0.988	0.668	0.000	0.985	0.981	0.730	0.032
envir					0.984	1.000	0.772	0.210
lowPFI= 0	0.977	0.998	0.693	0.166				
lowPFI= 1	0.987	0.999	0.796	0.255				
protest					0.879	0.668	0.360	0.065
lowPFI= 0	0.831	0.562	0.330	0.035				
lowPFI= 1	0.885	0.666	0.434	0.054				

Table 3.4: Item response probabilities for the four lower-level (individual-level) classes, describing different profiles of engaged citizenship norms. The probabilities are shown for a model where the measurement models of items *envir* and *protest* are non-equivalent with respect to the binary covariate *lowPFI* (countries with high vs. low levels of press freedom), and for a model where all the measurement probabilities are equivalent across countries.

Table 3.5 shows the estimated proportions of the latent classes after this first step, again for the selected model and for the full equivalence model for comparison. For the selected model, these probabilities are averaged over the sample proportions of the two values of *lowPFI*. In broad terms, the most noticeable difference between the higher-level classes is that one of them (class 2 in the table) has substantially higher probabilities than the other two classes of individuals belonging to the two lower-level classes (class *Medium* and *Unengaged*) which place least importance on these items as indicators of good citizenship. Averaged over the probabilities of the higher-level classes, the estimated proportions of individuals in the lower-level classes in the selected model are 0.37, 0.26, 0.29 and 0.07 for the *Maximal*, *Topic*, *Medium* and *Unengaged* classes respectively.

	Selected model Higher-level class (proportion)			Model with full measurement equivalence Higher-level class (proportion)		
Lower-level class	1 (0.500)	2 (0.375)	3 (0.125)	1 (0.500)	2 (0.375)	3 (0.125)
Maximal	0.467	0.290	0.236	0.428	0.189	0.289
Topic	0.251	0.190	0.543	0.233	0.151	0.549
Medium	0.226	0.408	0.185	0.273	0.493	0.141
Unengaged	0.055	0.112	0.036	0.066	0.167	0.021

Table 3.5: Estimated proportions of the three higher-level (country-level) latent classes, and of the four lower-level (individual-level) latent classes within the higher-level classes. The probabilities are shown for the selected model where the measurement models of items *envir* and *protest* are non-equivalent with respect to the binary covariate *lowPFI* (and averaging over the sample distribution of this variable) and for a model where all the measurement probabilities are equivalent across countries.

Finally, we estimate the structural model for the individual-level latent class given the covariates *lowPFI* and *lnGDPusd*. The estimated coefficients of this multinomial logistic model are reported in Table 3.6, again showing results based on the selected measurement model with MNE and, for comparison, a model with full measurement equivalence. Table 3.6 shows two-step estimates of the parameters of the structural model, estimated as described in Section 3.3, and with the measurement parameters fixed at their estimated values from Table 3.4. The reference category for a respondent is here the class *Unengaged*. Considering the estimates from the selected model, the results show that adolescents living in countries with less press freedom are increasingly more likely to have norms that emphasize more activities, relative to having "unengaged" norms, even after controlling for GDP. The coefficients of *lnGDPusd* indicate that individuals in higher-GDP countries are most likely to belong to the class *Maximal* which regards all the activities as important for good citizenship, but less likely to belong to the class *Topic* which de-emphasises the role of discussion only and (to a lesser extent) participation in peaceful protests. These differences between the more engaged classes (*Maximal* and *Topic*) and the less engaged classes (*Medium* and *Unengaged*) are substantively large and statistically significant (by conventional criteria) with respect to both covariates. In contrast, neither covariate makes a significant difference on the distinction between the two less engaged classes.

	Coefficient (in model vs. class *Unengaged*)					
	Selected model			Model with full measurement equivalence		
Covariate	*Maximal*	*Topic*	*Medium*	*Maximal*	*Topic*	*Medium*
lowPFI	0.963^{***}	1.657^{***}	0.423	1.125^{***}	1.987^{***}	0.619^{***}
lnGDPusd	0.561^{***}	-0.371^{**}	0.184	0.516^{*}	-0.440^{*}	0.059

Table 3.6: Estimated coefficients of the covariates *lowPFI* (dummy variable for countries that have low Press Freedom Index) and *lnGDPusd* (country's log GDP in US dollars) in a multilevel model for individual-level latent classes. These estimates are from the second step of two-step estimation. The measurement model for the items given the latent classes is fixed at the estimated parameters of the selected model which allows for measurement non-equivalence in two items (on the left) or of a model where all the measurement probabilities are equivalent across countries (on the right). The fixed measurement probabilities of these two choices are as shown in Table 3.4.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Comparing the estimates under the two specifications on measurement, we can see, in particular, that the coefficients of *lowPFI* are consistently less strong when we allow for measurement non-equivalence. This happens because some of the association between *lowPFI* and the responses is accounted for by measurement differences, that is, by the fact that the specific activities that have MNE are overall relatively more salient for adolescents in countries with low press freedom. where *lowPFI* is 1. Even after accounting for this, however, it is clear that adolescents in countries with low levels of press freedom are substantially more likely to be of the view that good citizenship is something that encompasses a larger number of activities.

3.6 Concluding remarks

We proposed a two-step estimation approach for multilevel latent class models with covariates in the presence of measurement non-equivalence. The method involves estimating the measurement model in the first step, and then holding its parameters fixed at their estimated values in the second step where the structural model for the classes given covariates is estimated. The key modification that is needed here, compared to two-step estimation of models with full measurement equivalence, is that covariates which create non-equivalence of measurement need to be included already in the first step. Their direct effects on measurement indicators are estimated there, while their coefficients (and those of any other covariates) in the structural model are estimated in the second step.

From a simulation study we observed that the proposed estimator performs generally well when the model is correctly specified, and essentially as well as the one-step maximum likelihood estimator which estimates all parameters at once. The performance of both estimators deteriorates to some extent in settings where the measurement model is weak and there is strong non-equivalence of measurement. The simulations also gave some evidence that two-step estimates are more robust to model misspecification which occurs when the measurement is incorrectly taken to be equivalent.

We have argued that two-step estimation has in principle two kinds of advantages over one-step estimation, the computational and the conceptual. The conceptual one is that estimating and fixing the measurement model before we proceed to estimate structural models for the latent classes means that the *definition* of the classes is then also fixed, and will not change even if we estimate and compare multiple different structural models. This advantage holds unchanged even when the models involve measurement non-equivalence. The computational advantage of two-step estimation, on the other hand, is somewhat reduced here. This is because the first step now includes also those covariates that are needed to account for the non-equivalence, making this step too more complex in comparison to when there is an absence of non-equivalence. It remains the case, however, that thereafter estimation is less demanding than it would be in one-step estimation. This is because in the two-step approach it will involve only the structural parameters, whereas the measurement parameters are fixed rather than repeatedly re-estimated.

As always, some questions on the properties and procedures of these methods are left open. We mention in particular questions of model selection. For multilevel latent class models with measurement non-equivalence this involves multiple dimensions: choosing the number of latent classes at the lower and higher levels, as well as determining which covariates are involved in non-equivalence and in what ways. Decisions on these dimensions could affect each other. In this paper we did not examine this question but employed a particular approach in line with previous literature. However, more systematic understanding of different approaches that could be used here would still be desirable.

Chapter 4

A Bayesian approach to two-step latent class analysis with covariates

Johan Lyrvall, Roberto Di Mari[1]*, & Jouni Kuha*[2]

4.1 Introduction

This study contributes to the methodological scholarship on two-step latent class analysis (LCA; first introduced by Lazarsfeld and Henry in 1968) by exploring the potential for Bayesian routines in improving statistical performance. In two-step LCA (Bakk & Kuha, 2018; Di Mari et al., 2023b; Lyrvall, Di Mari, Bakk, Oser, & Kuha, 2025), the model parameters are fitted in two separate steps. Step 1 involves fitting a *measurement model* to identify a set of categorical *latent classes* for some typology of units based on multiple categorical observed indicators, or *items.* The subsequent step 2 involves fitting a structural model for the latent classes given some observed covariates. The substantive rationale for separating these estimations into different steps lies in the avoidance of interpretational confounding. This can otherwise occur when the full parameter space is fitted simultaenously and the inclusion of the covariates distorts the measurement model defining the latent classes.

The two-step approach has been shown to be statistically suitable when the sample size is large and the latent classes well distinguishable from each other. In contrast, when the sample is small and class separation weak, two-step estimates of the structural model tend to be biased (Bakk & Kuha, 2018; Lyrvall et al., 2024). Methodological developments of two-step LCA have been focusing on the frequentist framework. Motivated by the general feature of Bayesian modeling to perform better than frequentist modeling in small samples, in this article we study an alternative two-step approach involving Bayesian estimation as a plausible method of improving the performance of the structural estimates in problematic modeling conditions. Different ways of integrating Bayesian routines into the two-step approach can be reasonably considered. Because the problematic performance of the conventional two-step approach arises when the measurement model tends to be more difficult to identify, a natural first investigation is to consider Bayesian estimation exclusively for the first step. As such, in this article we consider an alternative approach which fits the measurement model by means of a Bayesian routine in step 1, and the structural model by means of the conventional frequentist routine in step 2.

We investigate the performance of the alternative Bayesian two-step approach for the standard, single-level LCA modeling framework and the multilevel framework. Multilevel LCA (Vermunt, 2003) is used to analyze hierarchical data like cross-country surveys. Hierarchical data often exhibit dependencies between the nested units (e.g. survey respondents) across the higher-level units (e.g. countries). To model these dependencies, multilevel LCA can be used to introduce a second latent

[1]Department of Economics and Business, University of Catania, Italy
[2]Department of Statistics, London School of Economics, United Kingdom

class variable on the higher level (resulting, for example, in a clustering of individuals within a clustering of countries). In the last decade, interest in multilevel LCA has been increasing among applied scholars.

Bayesian estimation routines for single-level latent class measurement models have been presented for the single-level context by Li, Lord-Bessen, Shiyko, and Loeb (2018); Qiu (2022); White and Murphy (2014). Bayesian LCA has been considered for specific data analytical tasks such as multiple imputation of missing data (Vidotto, Vermunt, & Van Deun, 2018), identification of residual dependencies between the item responses (Lee, Jung, & Park, 2020), and modeling of within-cluster associations between the item responses (Malsiner-Walli, Grün, & Frühwirth-Schnatter, 2025). Bayesian modeling in two-step latent variable modeling has been previously considered by (Kuha, Zhang, & Steele, 2023; Zhang, Kuha, & Steele, 2024), whose main interests lie in continuous latent variables in a single-level context. These approaches estimate the step-1 model within the frequentist framework, and the step-2 model within the Bayesian framework. The approaches are developed within the context of particular applied analyses, with an absence of comparison of any alternative approach. The present study contributes to the latent variable modeling methodological literature with the first study of Bayesian estimation in two-step single-level and multilevel latent class analysis, including comparison of this approach with the conventional frequentist approach.

The paper is organized as follows. The next section describes the standard and multilevel frameworks for LCA with covariates. The following section describes the conventional frequentist two-step approach and presents the alternative Bayesian two-step approach. Next, we compare the performance of the Bayesian two-step approach to the frequentist two-step approach by means of a simulation study. We then illustrate the Bayesian approach by means of empirical examples. The final section discusses avenues for future research.

4.2 Modeling framework

We denote by X the latent class variable, $t = 1, \ldots, T$ its T discrete latent classes, $Y_1, \ldots, Y_H$ its H discrete indicators (items) with corresponding observed responses $y_{r1}, \ldots, y_{rH}$, and $\mathbf{Z}$ its predictors (covariates). We consider the following standard, single-level latent class model with covariates:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_H = y_H|\mathbf{Z}) &= \sum_{t=1}^{T} P(X = t|\mathbf{Z}) P(Y_1 = y_1, \ldots, Y_H = y_H|X = t) \\ &= \sum_{t=1}^{T} P(X = t|\mathbf{Z}) \prod_{h=1}^{H} P(Y_h = y_h|X = t), \end{aligned} \tag{4.1}$$

which takes the items to be independent of the covariates given the latent class variable (the *measurement equivalence* assumption), and the items to be independent of each other given the latent class variable (the *local independence* assumption). The $P(Y_h = y_h|X = t)$ define the latent class *measurement model*, and the $P(X = t|\mathbf{Z})$ the *structural model*. We take the structural model to be parameterized by multinomial logistic coefficients, such that

$$P(X = t|\mathbf{Z}) = \frac{\exp(\gamma_t'\mathbf{Z})}{1 + \sum_{s=2}^{T} \exp(\gamma_s'\mathbf{Z})}, \tag{4.2}$$

where $\gamma_1 = \mathbf{0}$ for identification. If one element in $\mathbf{Z}$ is a constant 1, the corresponding element in γ_t is an intercept.

For the multilevel modeling context (Vermunt, 2003), we denote by W the higher-level latent class variable, and $m = 1, \ldots, M$ its M discrete latent classes. For example, X may be defined on the individual level and W on the national level. We consider the following multilevel latent class model with covariates:

$$P(Y_1 = y_1, \dots, Y_H = y_h|\mathbf{Z}) = \sum_{m=1}^{M} P(W = m) \sum_{t=1}^{T} P(X = t|W = m, \mathbf{Z}) \prod_{h=1}^{H} P(Y_h = y_h|X = t), \tag{4.3}$$

which assumes measurement equivalence, local independence, and independence between W and $\mathbf{Z}$ given X. We take the corresponding structural model to be the multinomial logistic function

$$P(X = t|W = m, \mathbf{Z}) = \frac{\exp(\gamma'_{tm}\mathbf{Z})}{1 + \sum_{s=2}^{T} \exp(\gamma'_{sm}\mathbf{Z})}, \tag{4.4}$$

where $\gamma_{1m} = \mathbf{0}$ for identification.

4.3 Two-step estimation

We consider a sample of respondents $i = 1, \dots, N$. In the multilevel modeling context, we consider these to be nested within higher-level units $j = 1, \dots, J$, where j contains n_j respondents.

4.3.1 The conventional frequentist two-step approach

Let β be the parameter space for the measurement model, and γ be the parameter space for the structural model. In the single-level context, γ is defined by (4.2); In the multilevel context, it is defined by (4.4). Let α be the parameter space for the higher-level class proportions, $P(W = m)$, in (4.3).

The conventional frequentist two-step estimation approach for fitting γ for single-level latent class models with covariates (Bakk & Kuha, 2018) involves the following two steps. In step 1, the latent class model without covariates is fitted by means of maximizing the log-likelihood function

$$\ell_1 = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X = t) \prod_{h=1}^{H} P(Y_h = y_{ih}|X = t) \right].$$

The resulting measurement model estimates $\widehat{\beta}$ are retained and carried forward to step 2. The estimates for the other model parameters are discarded. In step 2, (4.1) is fitted conditional on $\widehat{\beta}$ by means of maximizing the pseudo log-likelihood function

$$\ell_2(\gamma|\beta = \widehat{\beta}) = \sum_{i=1}^{N} \log \left[\sum_{t=1}^{T} P(X = t|\mathbf{Z}_i) \prod_{h=1}^{H} P(Y_h = y_{ih}|X = t) \right].$$

The resulting structural estimates $\widehat{\gamma}$ give the final two-step estimates for (4.2).

The extension of the two-step approach to multilevel latent class models (Di Mari et al., 2023b) is straightforward. This version of the estimator considers the step-1 log-likelihood function

$$\ell_1 = \sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W = m) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X = t|W = m) \prod_{h=1}^{H} P(Y_h = y_{ijh}|X = t) \right].$$

The resulting $\widehat{\beta}$ are carried forward to step 2, and the estimates for the other model parameters discarded. Step 2 of the multilevel extension of the two-step approach considers the pseudo log-likelihood function

$$\ell_2(\alpha, \gamma|\beta = \widehat{\beta}) =$$
$$\sum_{j=1}^{J} \log \left[\sum_{m=1}^{M} P(W = m) \prod_{i=1}^{n_j} \sum_{t=1}^{T} P(X = t|W = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_h = y_{ijh}|X = t) \right].$$

The resulting $\widehat{\gamma}$ give the final two-step estimates for (4.4).

The maximization of the log-likelihood functions and pseudo log-likelihood functions for the single-level and multilevel modeling contexts are typically carried out by means of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). This integrates the Newton-Raphson (NR) algorithm in step 2 to fit (4.2). Specialized latent class analysis software like the proprietary option Latent GOLD (Vermunt & Magidson, 2021b) and the `R` package `multilevLCA` (Lyrvall, Di Mari, et al., 2025) implement these estimation routines.

4.3.2 An alternative Bayesian two-step approach

We here present an alternative Bayesian two-step approach, in which the step-1 model is fitted within the Bayesian framework, and the step-2 model fitted within the frequentist framework of the conventional two-step approach. We focus the presentation on step 1.

This alternative approach fits the step-1 model by means of Gibbs sampling (Geman & Geman, 1984), a Markov chain Monte Carlo method. A Gibbs sampling routine for single-level latent class models without covariates has been presented outside the context of two-step estimation by White and Murphy (2014) and Li et al. (2018). Our presentation of step 1 for the single-level context draws on these authors. Our presentation of step 1 for the multilevel context is an original contribution.

For simplicity of presentation, let Y_h be binary, such that $y_h \in \{0, 1\}$. For simplicity of notation, let $\phi_{h|t} = P(Y_h = 1|X = t)$. We take the model parameters to follow conjugate prior distributions. Specifically, we take the prior distribution of $\phi_{h|t}$ to be the Beta distribution

$$P(\phi_{h|t}|\kappa_{h|t}, \lambda_{h|t}) \propto \phi_{h|t}^{\kappa_{h|t}-1}(1 - \phi_{h|t})^{\lambda_{h|t}-1},$$

where $\kappa_{h|t}$ and $\lambda_{h|t}$ are hyperparameters. The use of default (uniform, flat) priors corresponds to setting $\kappa_{h|t} = \lambda_{h|t} = 1 \; \forall \; h, t$.

For the single-level context, let, for simplicity of notation, $\pi_t = P(X = t)$, and $\pi = (\pi_1, \ldots, \pi_T)$. We take the prior distribution of π to be the Dirichlet distribution

$$P(\pi|\delta_1, \ldots, \delta_T) \propto \prod_{t=1}^{T} \pi_t^{\delta_t - 1},$$

where $\delta_1, \ldots, \delta_T$ are hyperparameters. The use of default priors corresponds to setting $\delta_t = 1 \; \forall \; t$. Let $\widehat{\psi}_{it} = P(X = t|Y_1 = y_{i1}, \ldots, Y_H = y_{iH})$ be the posterior class membership probability of t for i. $\widehat{\psi}_{it}$ is defined by the Bayes rule as follows:

$$\widehat{\psi}_{it} = \frac{P(X = t) \prod_{h=1}^{H} P(Y_h = y_{ih}|X = t)}{\sum_{t=1}^{T} P(X = t) \prod_{h=1}^{H} P(Y_h = y_{ih}|X = t)}$$

Before the first iteration of the Gibbs sampling run for the single-level context, a random sample of π and $\phi_{h|t}$ is drawn from the Dirichlet distribution and the Beta distribution. We can write these draws as

$$\pi^{(0)} \sim \text{Dirichlet}(\delta_1, \ldots, \delta_T) \tag{4.5}$$

$$\phi_{h|t}^{(0)} \sim \text{Beta}(\kappa_{h|t}, \lambda_{h|t}). \tag{4.6}$$

Each iteration r of the Gibbs sampling run involves drawing random samples of class membership and model parameters based on the previous iteration $r-1$. For $r=1$, the previous iteration is given by (4.5) and (4.6). The first sub-routine of r involves drawing a sample of class membership $\widehat{X}_i^{(r)}$ from the multinomial distribution based on $\widehat{\psi}_{it}$, that is,

$$\widehat{X}_i^{(r)} \sim \text{Multinomial}(1, \widehat{\psi}_{i1}^{(r-1)}, \ldots, \widehat{\psi}_{iT}^{(r-1)})$$

The subsequent sub-routine involves sampling class membership probabilities and conditional response probabilities based on the sample of class membership. Let $I(\widehat{X}_i^{(r)} = t)$ be equal to 1 if i was assigned to t in r, and 0 otherwise. This subsequent sub-routine can be written as

$$\pi^{(r)} \sim \text{Dirichlet}(\sum_{i=1}^{N} I(\widehat{X}_i^{(r)} = 1) + \delta_1, \ldots, \sum_{i=1}^{N} I(\widehat{X}_i^{(r)} = T) + \delta_T)$$

$$\phi_{h|t}^{(r)} \sim \text{Beta}(\sum_{i=1}^{N} y_{ih} I(\widehat{X}_i^{(r)} = t) + \kappa_{h|t}, \sum_{i=1}^{N} (1-y_{ih}) I(\widehat{X}_i^{(r)} = t) + \lambda_{h|t}). \quad (4.7)$$

The parameters $\sum_{i=1}^{N} I(\widehat{X}_i^{(r)} = t)$ and $\sum_{i=1}^{N} y_{ih} I(\widehat{X}_i^{(r)} = t)$ for these probability distributions are similarly computed in the EM algorithm for frequentist estimation of the same model. (4.7) shows how the empirical information summarized by these parameters is combined with the priors on the basis of the hyperparameters δ_t, $\kappa_{h|t}$, and $\lambda_{h|t}$ in the role of a smoothing quantities. After some maximum number of iterations R, the average response probability $\frac{1}{R}\sum_{r=1}^{R} \phi_{h|t}^{(r)}$ is computed for all h and t. This set of averages is retained as the final step-1 estimate $\widehat{\beta}$ and carried forward to step 2. The sets of $\widehat{X}_i^{(r)}$ and $\pi^{(r)}$ are discarded.

For the multilevel context, let $\omega_m = P(W=m)$, and $\omega = (\omega_1, \ldots, \omega_M)$. Let $\pi_{t|m} = P(X = t|W=m)$, and $\pi_m = (\pi_{1|m}, \ldots, \pi_{T|m})$. We take the prior distributions of ω and π_m to be the Dirichlet distributions

$$P(\omega|\rho_1, \ldots, \rho_M) \propto \prod_{m=1}^{M} \omega_m^{\rho_m - 1}$$

$$P(\pi_m|\delta_{1|m}, \ldots, \delta_{T|m}) \propto \prod_{t=1}^{T} \pi_{t|m}^{\delta_{t|m}-1},$$

where $\rho_1, \ldots, \rho_M$ and $\delta_{1m}, \ldots, \delta_{Tm}$ are hyperparameters. The use of default priors corresponds to setting $\rho_m = \delta_{tm} = 1 \; \forall \; t, m$. Let $\widehat{\mu}_{jm}$ be the posterior higher-level class membership probability of m for j. This is given by

$$\widehat{\mu}_{jm} = \frac{P(W=m)\prod_{i=1}^{n_j}\sum_{t=1}^{T} P(X=t|W=m)\prod_{h=1}^{H} P(Y_h = y_{ijh}|X=t)}{\sum_{m=1}^{M} P(W=m)\prod_{i=1}^{n_j}\sum_{t=1}^{T} P(X=t|W=m)\prod_{h=1}^{H} P(Y_h = y_{ijh}|X=t)}.$$

In the multilevel context, the lower-level class membership probability $\widehat{\eta}_{ijt}$ is given by

$$\widehat{\eta}_{ijt} = \sum_{m=1}^{M} \widehat{\mu}_{jm} \frac{P(X=t|W=m)\prod_{h=1}^{H} P(Y_h = y_{ih}|X=t)}{\sum_{t=1}^{T} P(X=t|W=m)\prod_{h=1}^{H} P(Y_h = y_{ih}|X=t)}.$$

Before the first iteration of the Gibbs sampling run for the multilevel context, a random sample is drawn for ω, π_m, and $\phi_{h|t}$. The draw for $\phi_{h|t}$ is given by (4.6). The draws for ω and π_m can be written as

$$\omega^{(0)} \sim \text{Dirichlet}(\rho_1, \dots, \rho_M)$$

$$\pi_m^{(0)} \sim \text{Dirichlet}(\delta_{1|m}, \dots, \delta_{T|m}).$$

The first sub-routine of r involves drawing a sample of higher-level class membership $\widehat{W}_j^{(r)}$ and lower-level class membership $\widehat{X}_i^{(r)}$, that is,

$$\widehat{W}_j^{(r)} \sim \text{Multinomial}(1, \widehat{\mu}_{j1}^{(r-1)}, \dots, \widehat{\mu}_{jM}^{(r-1)})$$

$$\widehat{X}_i^{(r)} \sim \text{Multinomial}(1, \widehat{\eta}_{ij1}^{(r-1)}, \dots, \widehat{\eta}_{ijT}^{(r-1)}).$$

The next sub-routine involves sampling higher-level class membership probabilities, lower-level class membership probabilities, and conditional response probabilities. Let $I(\widehat{W}_j^{(r)} = m)$ be equal to 1 if j was assigned to m in r, and 0 otherwise. The sampling for the conditional response probabilities is given by (4.7). We can write this next sub-routine as

$$\omega^{(r)} \sim \text{Dirichlet}(\sum_{j=1}^{J} I(\widehat{W}_j^{(r)} = 1) + \rho_1, \dots, \sum_{j=1}^{J} I(\widehat{W}_j^{(r)} = M) + \rho_M)$$

$$\pi_m^{(r)} \sim$$
$$\text{Dirichlet}(\sum_{j:\widehat{W}_j^{(r)}=m} \sum_{i=1}^{n_j} I(\widehat{X}_i^{(r)} = 1) + \delta_{1|m}, \dots, \sum_{j:\widehat{W}_j^{(r)}=m} \sum_{i=1}^{n_j} I(\widehat{X}_i^{(r)} = T) + \delta_{T|m}).$$

After R iterations, $\widehat{W}_j^{(r)}$, $\widehat{X}_i^{(r)}$, $\omega^{(r)}$, and $\pi_m^{(r)}$ are discarded, and the average response probability $\frac{1}{R}\sum_{r=1}^{R} \phi_{h|t}^{(r)}$ is computed for all h and t. The set of average response probabilities is retained as the final step-1 estimate $\widehat{\beta}$ and carried forward to step 2.

A distinct feature of Gibbs sampling in relation to the EM algorithm is the problem of label switching. One method of dealing with label switching, which is applied in the subsequent simulation study and empirical examples, was proposed for the single-level context by White and Murphy (2014) drawing on Wyse and Friel (2012). This method minimizes a cost function based on the number of matching class memberships between the current iteration and the first iteration. Let $\mathbf{C}^{(rp)}$ be a $T \times T$ cost matrix for iteration r and permutation p of the class labels, with elements $C_{t,s}^{(rp)} = \sum_{i=1}^{N} I(\widehat{X}_i^{(1)} = t) * I(\widehat{X}_i^{(r)} = s)$. The p for which the sum of diagonal entries is the largest, $\arg\max_p \sum_{t=1}^{T} C_{t,t}^{(rp)}$, is retained as the optimal permutation for r. In the multilevel context, we apply the analogous procedure, in which $\mathbf{C}^{(rp)}$ is a $M \times M$ cost matrix with elements $C_{m,l}^{(rp)} = \sum_{j=1}^{J} I(\widehat{W}_j^{(1)} = m) * I(\widehat{W}_j^{(r)} = l)$, and $\arg\max_p \sum_{m=1}^{M} C_{m,m}^{(rp)}$, is retained as the optimal permutation of the higher-level class labels for r.

Application of the Gibbs sampling routine often involves conventional convergence diagnostics for the number of burn-in iterations. There are different types of convergence diagnostics, such as visual inspection of diagnostics plots and quantitative measures such as the Raftery and Lewis diagnostic (Raftery & Lewis, 1992). This has been discussed by White and Murphy (2014) in the context of Bayesian LCA without covariates. In the present context of two-step Bayesian LCA with covariates, we consider an alternative strategy focusing on the two-step structural estimates. This involves evaluating, across considered set of candidate numbers of burn-in iterations, when the structural estimates stabilize to the same distribution. We illustrate this approach in the subsequent simulation study.

4.4 Simulation study

We compare the performance of the alternative Bayesian two-step approach with the conventional frequentist two-step approach by means of a simulation study. The focus lies in bias and efficiency. The latter is measured by means of the root mean squared error (RMSE). To apply the frequentist approach we use the `R` package `multilevLCA` (Lyrvall, Di Mari, et al., 2025). To apply the Bayesian approach we develop `C++` code integrated to `R` using the `RcppArmadillo` package (Eddelbuettel & Sanderson, 2014).

4.4.1 Design

The performance of the frequentist two-step approach in the single-level context (Bakk & Kuha, 2018) and the multilevel context (Di Mari et al., 2023b; Lyrvall et al., 2024) is known to depend on sample size and class separation. The performance tends to be better when the sample size is larger or the separation between the classes is stronger. For the single-level and multilevel contexts respectively we consider 6 fully-crossed modeling conditions of 3 different sample sizes and 2 different degrees of class separation. For the sample size we consider quantities of 500, 1000, and 2000, and refer to these as smaller, moderate, and larger sample sizes, respectively. In the LCA methodological literature, 500 is typically used as a minimal sample size in simulation studies to analyze small-sample properties of stepwise estimators (Asparouhov & Muthén, 2014; Bakk & Kuha, 2018; Vermunt, 2010).

In all modeling conditions the population model has 3 (lower-level) latent classes and 6 binary indicators. The multilevel model has 20 higher-level units nested within 2 higher-level latent classes with equal proportions. We vary the separation of the (lower-level) classes by means of their probability of scoring 1 on each of the indicators. Class 1 has high probabilities for all 6 indicators, class 2 has high probabilities for indicators 1-3 and low probabilities for indicators 4-6, and class 3 has low probabilities for all 6 indicators. We vary the degree of class separation by means of the exact quantity of these "high" and "low" probabilities. In the weaker-separation condition, the high probabilities are 0.75 and the low probabilities 0.25. In the stronger-separation condition, the high probabilities are 0.90 and the low probabilities 0.10.

We consider a single binary covariate with equal probabilities for value 1 and value 0. The reference class in the multinomial logistic models is (lower-level) class 1. The covariate effect on (lower-level) class 2 and class 3 is 1 in all specifications. Averaged over the distribution of the covariate, the classes in the single-level model have equal proportions. Averaged over the distribution of the covariate, the lower-level classes in the multilevel model have proportions $P(X = 1|W = 1) = P(X = 3|W = 2) = 0.18$, $P(X = 2|W = 1) = P(X = 2|W = 2) = 0.31$, and $P(X = 3|W = 1) = P(X = 1|W = 2) = 0.51$.

For each modeling condition, we apply the frequentist and Bayesian two-step approaches to 500 simulated random samples based on a correctly specified single-level or multilevel model. Samples yielding inadmissable values for either of the approaches are replaced. For each modeling condition, we evaluate bias and RMSE with respect to all the slope parameter estimates across the 500 samples. In the multilevel context, whereas the population model has fixed slopes across the higher-level classes, we estimate specifications in which the slope parameter is allowed to vary across the higher-level classes. As such, in the single-level and multilevel contexts, for each modeling condition, bias and RMSE are evaluated across $2 * 500 = 1000$ and $4 * 500 = 2000$ parameter estimates, respectively.

The Gibbs sampler is applied based on 750 retained iterations, a thinning rate of 1, and default priors. To select the number of burn-in iterations, we compare the bias for the most problematic of the considered modeling conditions, that is, the modeling condition combining the smallest sample size with the weakest class separation, across 3 candidates, based on 500 simulated random samples and a correctly specified model. The candidates are 250, 500, or 750 burn-in iterations. We perform this selection separately for the single-level context and the multilevel context based on visual inspection of the distribution of the structural estimates. The results are shown in Figure 4.1. As can be seen, for both the single-level context and the multilevel context, the resulting distributions are approximately stable across all these candidates, supporting the candidate of 250 burn-in iterations. The results for the simulation study presented in the next sub-section are based on this quantity.

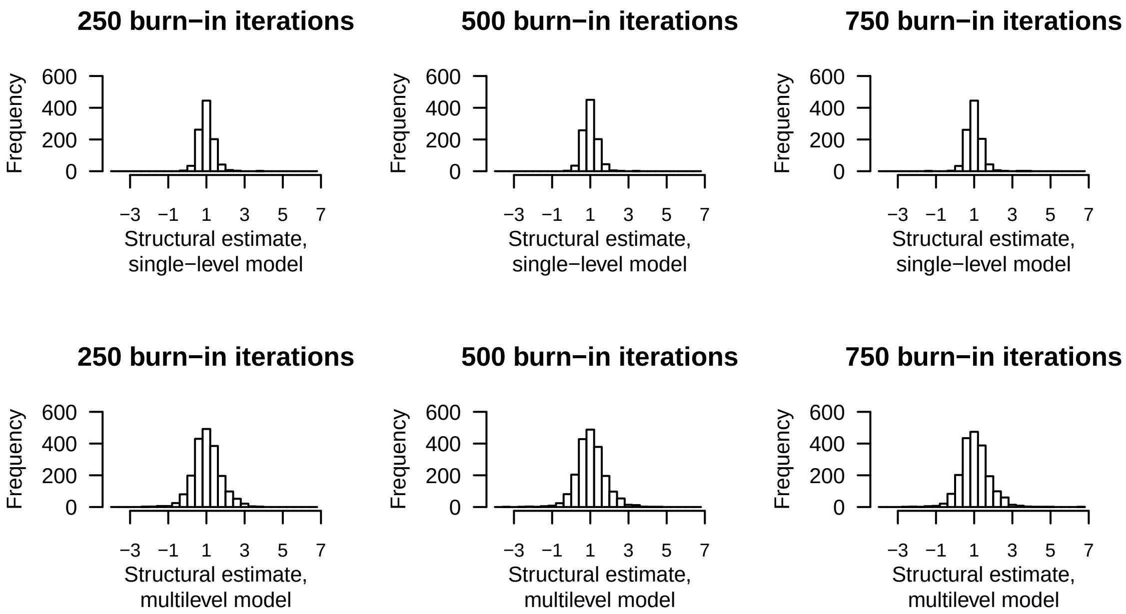

Figure 4.1: Distribution of structural estimates, across 500 simulated random samples, for the single-level model and the multilevel model with varying numbers of burn-in iterations.

4.4.2 Results

Table 4.1 reports the results for the single-level context. These results show the conventional frequentist two-step approach to be less biased and more efficient when the sample size is larger and class separation stronger. This is consistent with previous research (Bakk & Kuha, 2018). The frequentist approach is essentially unbiased when the weaker-separation condition is combined with the larger-sample-size condition, and in all the stronger-separation conditions. The alternative Bayesian two-step approach is essentially unbiased in all the considered modeling conditions.

As shown in Table 4.1, the Bayesian approach exhibits greater efficiency than the frequentist approach when the weaker-separation condition is combined with the smaller- and moderate-sample-size conditions. The frequentist approach and the Bayesian approach are as good as equally efficient when the weaker-separation condition is combined with the laer-sample size condition and in all the stronger-separation conditions.

Separation	Sample size	Mean bias (True value: 1)		RMSE (True value: 1)	
		Frequentist	Bayesian	Frequentist	Bayesian
Weaker	500	-0.10	0.01	0.40	0.33
	1000	-0.05	-0.01	0.27	0.23
	2000	-0.01	0.00	0.17	0.16
Stronger	500	0.00	0.00	0.24	0.23
	1000	0.00	0.00	0.17	0.17
	2000	-0.01	0.00	0.11	0.11

Table 4.1: Mean bias and root mean squared error (RMSE), across 500 simulated random samples, for single-level models with varying sample sizes and degrees of class separation for two-step estimates of slope parameters with population value 1, based on the conventional frequentist approach and the alternative Bayesian approach.

The results for the multilevel context are reported in Table 4.2. The observed pattern for the conventional frequentist two-step approach for multilevel models is consistent with previous research (Di Mari et al., 2023b; Lyrvall et al., 2024). Moreover, the relative performance of the frequentist approach and the Bayesian approach is comparable with the results for the single-level context. The frequentist approach is as good as unbiased in the moderate- and larger-sample-size conditions, and the Bayesian approach as good as unbiased in all the considered modeling conditions.

Separation	Sample size	Mean bias (True value: 1)		RMSE (True value: 1)	
		Frequentist	Bayesian	Frequentist	Bayesian
Weaker	500	0.10	0.03	1.21	0.71
	1000	0.01	0.01	0.41	0.41
	2000	-0.01	-0.01	0.28	0.28
Stronger	500	0.03	0.02	0.51	0.44
	1000	0.01	0.01	0.27	0.27
	2000	-0.01	-0.01	0.19	0.19

Table 4.2: Mean bias and root mean squared error (RMSE), across 500 simulated random samples, for multilevel models with varying sample sizes and degrees of class separation for two-step estimates of slope parameters with population value 1, based on the conventional frequentist approach and the alternative Bayesian approach.

As can be seen in Table 4.2, with respect to efficiency, the Bayesian approach outperforms the frequentist approach in the smaller-sample-size conditions. In the moderate- and larger-sample-size conditions, the frequentist approach and the Bayesian approach are essentially equally efficient. The frequentist approach and the Bayesian approach are consistently less efficient in the multilevel context compared to the single-level context.

The main difference in the results across the considered modeling conditions between the single-level context and the multilevel context can be observed when the the moderate-sample-size condition and the weaker-separation condition are combined. In this modeling condition, the frequentist approach yields some bias in the single-level context and essentially no bias in the multilevel context. In the same modeling condition, the Bayesian approach is more efficient than the frequentist approach in the single-level context, whereas the Bayesian approach and the frequentist approach are essentially equally efficient in the multilevel context.

4.5 Empirical examples

We illustrate the alternative Bayesian two-step approach in the applied single-level and multilevel contexts by means of two empirical examples. For comparison, we also apply the conventional frequentist two-step approach. As in the simulation study, we apply the Gibbs sampler based on 250 burn-in iterations, a thinning rate of 1, and default priors. Because the correct specification of the model is unknown, we apply a larger number of retained iterations than in the simulation study, namely 3000 retained iterations. The illustration for the single-level context is an analysis of jazz artists' professional recognition, and the illustration for the multilevel context an analysis of cultural participation types.

4.5.1 Jazz artists' professional recognition

In this empirical example, we analyze professional recognition as a potential form of psychic income in labor markets for the creative industries. Psychic income is derived from the intrinsic pleasure of artistic production. The applied cultural economics literature has put forth theoretical and empirical arguments that artists typically consider psychic incomes as compensating for low monetary incomes in their labor supply decisions (Longden & Throsby, 2021; Throsby, 1994). We focus on jazz musicians in the United States, using survey data from the *Study of Jazz Artists 2001* (Jeffri, 2015). This survey effort applied a respondent-driven sampling design (the sampling design for this particular survey effort is documented in Heckathorn & Jeffri, 2001).

The full sample contains 674 jazz musicians. We focus on the 597 of these that stated that their talent had been recognized. The survey data contain binary measures of 6 specific kinds of professional recognition, documenting whether the respondent's talent has been recognized from performing widely (*performing*), radio coverage (*radio*), newspaper articles (*newspaper*), magazine articles (*magazine*), TV coverage (*tv*), or recording with a major record label (*recording*). The latent class analysis in this sub-section combines these 6 items into distinct patterns of professional recognition.

As covariates we consider 2 binary variables: whether the most important factor prompting the respondent's initial pursuit of a career in music was an inner drive to make music (*drive*), and whether the respondent earns more than half of his or her income from jazz activities (*income*). For expositional efficiency, we explain the substantive relevance of the covariatess in the interpretation of the results. In the second step of the two-step approach, rows with missing values on the covariates are discarded, leading to a sample size of 567.

Before reporting the point estimates for this model, we investigate potential label switching based on the diagnostics plot in Figure 4.2. This shows the response probabilities for each class for each of the 3000 retained iterations. As can be seen, label switching is corrected for succesfully.

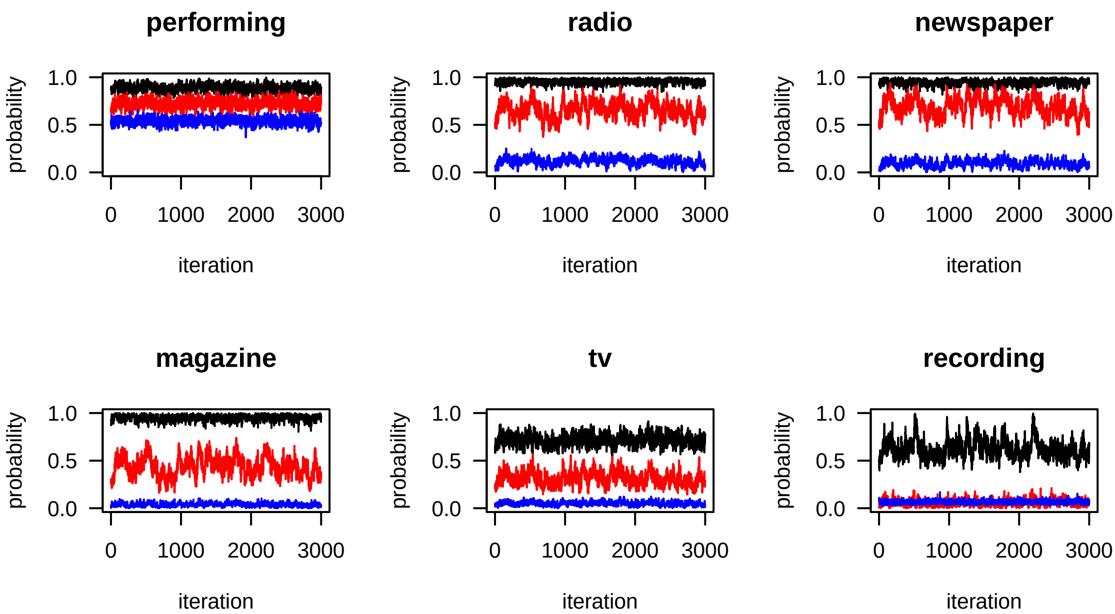

Figure 4.2: Diagnostics plot of the response probabilities for the 3000 retained iterations of the Gibbs sampling run for the single-level model with 3 latent classes of jazz artists' professional recognition. Class 1 is described in black, class 2 in red, and class 3 in blue.

Table 4.3 reports the two-step point estimates of the latent class model of patterns of professional recognition of jazz musicians. The substantive interpretations of the estimated latent classes are effectively the same for the frequentist approach and the Bayesian approach. As can be seen in the results for the measurement model, class 1 has high probabilities of self-reported recognition across all 6 items. Class 2 has high probabilities of recognition for the items that typically have a local outreach, namely performing widely, radio coverage, and newspaper articles; and low probabilities of recognition for those that often have a larger geographical outreach, namely magazine articles, TV coverage, and recording with a major record label. Class 3 has high recognition probability for performing widely, and low recognition probabilities for the remaining 5 items. We can label class

1, class 2, and class 3 as the *widely recognized*, *moderately recognized*, and *narrowly recognized* types, respectively.

	Frequentist			Bayesian		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Measurement model						
performing	0.90	0.73	0.54	0.89	0.72	0.54
radio	0.97	0.67	0.11	0.96	0.66	0.12
newspaper	0.97	0.70	0.09	0.95	0.69	0.09
magazine	0.98	0.45	0.03	0.95	0.44	0.04
tv	0.74	0.31	0.04	0.72	0.31	0.05
recording	0.64	0.06	0.06	0.63	0.05	0.06
Structural model						
drive		0.76	0.89		0.75	0.86
income		-1.26	-2.42		-1.21	-2.36
intercept		1.25	2.13		1.10	2.04

Table 4.3: Fitted values for the single-level latent class model with covariates, based on the conventional frequentist two-step approach and the alternative Bayesian two-step approach.

The structural-model point estimates reported in Table 4.3 exhibit similar patterns for the frequentist approach and the Bayesian approach. The jazz artists that initially pursued a career in music due to an inner drive to make music are markedly more likely than those with other initial influences to belong to the moderately recognized type relative to the widely recognized type, and still more likely to belong to the narrowly recognized type relative to the widely recognized type. This finding suggests that realizing an inner drive may function as a compensating form of psychic income compensating for a lack of professional recognition. The jazz artists that derive most of their income from jazz activities are substantially more likely than those who earn most of their income from different sources to belong to increasingly less recognized types compared to the widely recognized type. This suggests that a higher marginal productivity of jazz artists may yield both substantively significant financial incomes and psychic incomes in the form of professional recognition.

4.5.2 Cultural participation types

This empirical example analyses types of cultural participants on the extensive margin across 6 activities. Similar studies have been conducted in the applied sociological and cultural-economic scholarships to investigate how individuals combine highbrow (high-status) and lowbrow (low-status) cultural activities into distinct strata (see, for example, Van Rees, Vermunt, & Verboord, 1999). We conduct a two-step multilevel latent class analysis using survey data from the *Local Area Arts Participation Study 1992* (National Endowment for the Arts, 2015), which applied random sampling to 12 cities and areas in the United States. The lower level and higher level in our analysis correspond to the respondent level and the city/area level, respectively.

The items for cultural participation measure whether, in the prior 12 months, the respondent attended a live jazz performance (*jazz*), a live classical music performance (*classic*), a live musical stage play or operetta (*musical*), a live non-musical stage play (*play*), a live ballet performance (*ballet*), or an art museum or gallery (*museum*). We consider a model with 3 lower-level classes and 2 higher-level classes. To investigate the relationship between cultural participation types and socioeconomic status, we consider as covariates continously coded measures of education (*education*) and income (*income*). The sample size for the first step of the two-step approach is 1420. Due to missing values in the covariates, the sample size for the second step is 1164.

The diagnostics plot in Figure 4.3 displays the response probabilities for each class for the 3000 retained iterations of the Gibbs sampling run. This shows that label switching is corrected for successfully.

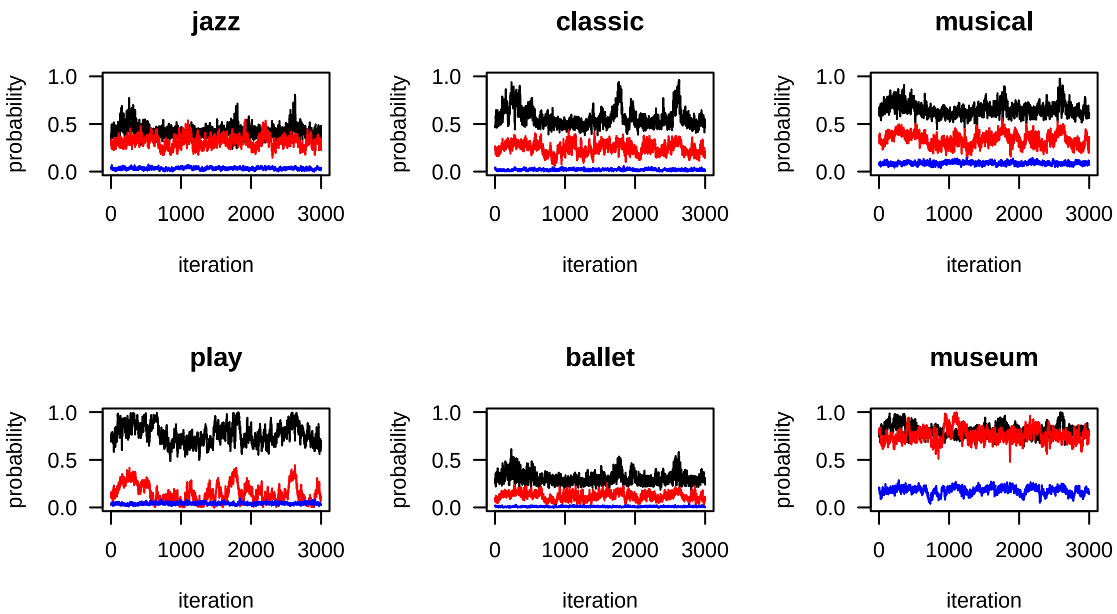

Figure 4.3: Diagnostics plot of the response probabilities for the the 3000 retained iterations of Gibbs sampling run for the multilevel model with 3 lower-level latent classes of cultural participation types. Class 1 is described in black, class 2 in red, and class 3 in blue.

The point estimates for the fitted multilevel latent class model with covariates is reported in Table 4.4. The frequentist approach and the Bayesian approach yield comparable class solutions. Based on the reported estimates for the measurement model, class 1, class 2, and class 3 can be labeled as the *omnivore* type, the *museum-only* type, and the *inactive* type, respectively. The omnivores have the largest probability of particiation for each of the 6 items. The museum-only type focuses his or her cultural participation on art museums or galleries. The inactive type has small probabilities of participation for each item.

The observed point estimates for the structural models in Table 4.4 are also comparable across the frequentist approach and the Bayesian approach. In both of the higher-level classes, education and income are negatively associated with the museum-only type and the inactive type. The magnitudes of these associations are stronger for the inactive type than for the museum-only type. These patterns suggest that individuals with increasingly broader repertoires of participation have increasingly higher socioeconomic status.

4.6 Concluding remarks

This study explored a plausible method of improving the performance of the state-of-the-art two-step estimation approach (Bakk & Kuha, 2018; Di Mari et al., 2023b; Lyrvall, Di Mari, et al., 2025), namely, integration of Bayesian estimation. The conventional frequentist two-step approach is known to be biased when the latent class measurement model is more difficult to identify due to small samples and weak class separation (Bakk & Kuha, 2018; Lyrvall et al., 2024). Bayesian modeling tends to perform better than frequentist modeling in small samples. We focused on an alternative two-step approach wherein the estimation of the measurement model replaced the EM algorithm with Gibbs sampling.

	Frequentist			Bayesian		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Measurement model						
jazz	0.39	0.33	0.03	0.42	0.32	0.03
classic	0.49	0.25	0.02	0.57	0.24	0.02
musical	0.62	0.32	0.09	0.65	0.34	0.09
play	0.83	0.01	0.04	0.77	0.15	0.04
ballet	0.26	0.13	0.00	0.31	0.12	0.01
museum	0.79	0.76	0.18	0.81	0.76	0.18
Structural model						
HL class 1						
education		-0.21	-0.84		-0.27	-0.90
income		-0.14	-0.22		-0.14	-0.24
intercept		2.01	6.09		2.86	6.84
HL class 2						
education		-0.01	-0.53		-0.07	-0.55
income		0.02	-0.33		0.00	-0.36
intercept		0.43	6.08		1.27	6.68

Table 4.4: Fitted values for the multilevel latent class model with covariates, based on the conventional frequentist two-step approach and the alternative Bayesian two-step approach.

We analyzed the alternative Bayesian two-step approach for standard single-level latent class models and for multilevel latent class models (Vermunt, 2003). In a simulation study, we observed that the Bayesian approach reduces the bias of the conventional approach in problematic modeling conditions. The Bayesian approach was also observed to be more efficient than the conventional approach in problematic modeling conditions. In more advantageous modeling conditions, the Bayesian approach and the conventional approach were observed to perform essentially equally with respect to unbiasedness and efficiency.

An issue that deserves further investigation is the computation of standard errors of the two-step estimates for the alternative Bayesian approach. For the conventional frequentist approach, this involves computing the variability arising directly from the step-2 estimator and the additional variability arising indirectly from the step-1 estimator. When the step-1 estimator is Bayesian, the rationale and mechanism of analogous correction is not obvious considering the conceptual differences between Bayesian statistics and frequentist statistics. Another issue deserving further research is the performance of the Bayesian two-step approach in handling missing data in the indicators, in light of the popularity of Bayesian statistics in general for this task because of its flexibility. Finally, yet another issue that deserves further investigation is the performance of the Bayesian approach as a plausible method for reducing bias in two-step latent class analysis in the presence of direct effects from the covariates to the indicators (differential item functioning, measurement non-invariance or non-equivalence). Recently, the conventional frequentist approach was extended to multilevel latent class models with direct effects by Lyrvall, Kuha, and Oser (2025). These authors observe a similar pattern of performance, with the additional dimension of problematic performance when the number of indicators subject to direct effects is large. Given the results of the present study, we can reasonably extrapolate that the Bayesian approach be a promising method for handling this kind of additional complexities and expand the usefulness of two-step latent class analysis in general to a wider range of modeling contexts.

Concluding remarks

Important limitations of this thesis and promising a include the following. The developments of stepwise estimation approaches for multilevel latent class models in the second, third, and fourth chapter, omit consideration of the computation of bias-corrected standard errors of the structural estimates. Methodological developments of methods to compute these quantities would further extend the applicability of the proposed approaches. Such studies are a promising avenue for future research. Furthermore, the analyses of the statistical properties of the stepwise approaches proposed in the second, third, and fourth chapter, omit investigating these properties when the model is not correctly specified. In applied LCA, the "true" model specification is typically unknown and must be selected. Such extended analyses are another promising avenue for future research. Yet another promising avenue for future multilevel LCA methodological contributions is the extension of `multilevLCA` to automatically implement methods developed in the second, third, and fourth chapter. Finally, an additional promising avenue for future research lies in developing stepwise estimation approaches combining the novelties of the those proposed in the second, third, and fourth chapter; for instance, bias-adjusted three-step estimation of multilevel models with measurement non-equivalence, and Bayesian two-step estimation of multilevel models with measurement non-equivalence.

References

Aitkin. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, *55*(1), 117–128.

Aitkin, & Alfó. (1998). Regression models for binary longitudinal responses. *Statistics and Computing*, *8*, 289–307.

Asparouhov, & Muthén. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341.

Bakk, Di Mari, Oser, & Kuha. (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, & Kuha. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*(4), 871–892.

Bakk, & Kuha. (2021). Relating latent class membership to external variables: An overview. *British Journal of Mathematical and Statistical Psychology*, *74*, 340–362.

Bakk, Oberski, & Vermunt. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*(4), 520–540.

Bakk, Tekle, & Vermunt. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311.

Bandeen-Roche, Miglioretti, Zeger, & Rathouz. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375–1386.

Bartolucci, Bacci, & Gnaldi. (2014). Multilcirt: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, *71*, 971–985.

Bartolucci, Montanari, & Pandolfi. (2016). Item selection by latent class-based methods: An application to nursing home evaluation. *Advances in Data Analysis and Classification*, *10*, 245–262.

Bijmolt, Paas, & Vermunt. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing*, *21*(4), 323–340.

Bolck, Croon, & Hagenaars. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27.

Bray, Watson, Salisbury-Afshar, Taylor, & McGuire. (2023). Patterns of opioid use behaviors among patients seen in the emergency department: Latent class analysis of baseline data from the point pragmatic trial. *Journal of Substance Use and Addiction Treatment*, *146*, 208979.

Collins, Fidler, Wugalter, & Long. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*(3), 375–389.

Dean, & Raftery. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, *62*, 11–35.

Dempster, Laird, & Rubin. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(1), 1–22.

Di Mari, & Bakk. (2018). Mostly harmless direct effects: A comparison of different latent Markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 467-483.

Di Mari, Bakk, Oser, & Kuha. (2023a). Multilevel latent class analysis with covariates: Analysis of cross-national citizenship norms with a two-stage approach. *arXiv preprint arXiv:2307.10720*.

Di Mari, Bakk, Oser, & Kuha. (2023b). A two-step estimator for multilevel latent class analysis with covariates. *Psychometrika*, *88*(4), 1144-1170.
Di Mari, & Lyrvall. (2024). multilevLCA: Estimates and plots single-level and multilevel latent class models [Computer software manual]. (R package)
Di Mari, Oberski, & Vermunt. (2016). Bias-adjusted three-step latent markov modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 649–660.
Eddelbuettel, & François. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.
Eddelbuettel, & Sanderson. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data analysis*, *71*, 1054–1063.
Fagginger Auer, Hickendorff, Van Putten, Béguin, & Heiser. (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*(2), 144–159.
Finch, & French. (2014). Multilevel latent class analysis: Parametric and nonparametric models. *The Journal of Experimental Education*, *82*(3), 307–333.
Geman, & Geman. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *??*(6), 721–741.
Gnaldi, Bacci, & Bartolucci. (2016). A multilevel finite mixture item response model to cluster examinees and schools. *Advances in Data Analysis and Classification*, *10*(1), 53–70.
Goodman. (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*(5), 1179–1259.
Goodman. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.
Goodman. (2007). On the assignment of individuals to latent classes. *Sociological Methodology*, *37*(1), 1–22.
Hagenaars. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage.
Hagenaars. (1992). *Exemplifying longitudinal log-linear analysis with latent variables.* European Science Foundation, Scientific Network on Household Panel Studies.
Hallquist, & Wiley. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.
Heckathorn, & Jeffri. (2001). Finding the beat: Using respondent-driven sampling to study jazz musicians. *Poetics*, *28*(4), 307–329.
Henry, & Muthén. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 193–215.
Hickendorff, van Putten, Verhelst, & Heiser. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology*, *102*(2), 438–452.
Hooghe, & Oser. (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, Oser, & Marien. (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.
Janssen, Van Laar, De Rooij, Kuha, & Bakk. (2019). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 280–290.
Jeffri. (2015). *Study of Jazz Artists 2001 [United States].* Inter-university Consortium for Political and Social Research [distributor]. Retrieved from `https://doi.org/10.3886/ICPSR35236.v1`
Kankaraš, Moors, & Vermunt. (2018). Testing for measurement invariance with latent class analysis. In *Cross-cultural analysis* (pp. 393–419). Routledge.
Kankaraš, Vermunt, & Moors. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods and*

Research, *40*, 279–310.
Kim, Jeon, Chang, & Chung. (2022). glca: An R package for multiple-group latent class analysis. *Applied Psychological Measurement*, *46*(5), 439–441.
Kuha, & Bakk. (2023). *Two-step estimation of latent trait models.* (arXiv preprint arXiv: 2303.16101)
Kuha, Zhang, & Steele. (2023). Latent variable models for multivariate dyadic data with zero inflation: Analysis of intergenerational exchanges of family support. *The Annals of Applied Statistics*, *17*(2), 1521–1542.
Lacourse, de la Sablonnière, Giguère, Morin, Legault, & Laliberté. (2024). stepmixr: Interface to Python package StepMix [Computer software manual]. (R package)
Laird. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.
Lanza, & Rhoades. (2013). Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*, *14*, 157–168.
Lanza, Tan, & Bray. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 1–26.
Lazarsfeld, & Henry. (1968). *Latent structure analysis.* Houghton Mifflin.
Lee, Jung, & Park. (2020). Detecting conditional dependence using flexible Bayesian latent class analysis. *Frontiers in Psychology*, *11*, 1987.
Li, Lord-Bessen, Shiyko, & Loeb. (2018). Bayesian latent class analysis tutorial. *Multivariate Behavioral Research*, *53*(3), 430–451.
Linzer, & Lewis. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29.
Longden, & Throsby. (2021). Non-pecuniary rewards, multiple job-holding and the labour supply of creative workers: the case of book authors. *Economic Record*, *97*(316), 24–44.
Lukočienė, Varriale, & Vermunt. (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, Bakk, Oser, & Di Mari. (2024). Bias-adjusted three-step multilevel latent class modeling with covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(4), 592–603.
Lyrvall, Di Mari, Bakk, Oser, & Kuha. (2023). multilevLCA: An R package for single-level and multilevel latent class analysis with covariates. *arXiv preprint arXiv:2305.07276*.
Lyrvall, Di Mari, Bakk, Oser, & Kuha. (2025). Multilevel latent class analysis: State-of-the-art methodologies and their implementation in the R package multilevLCA. *Multivariate Behavioral Research*, 1–17.
Lyrvall, Kuha, & Oser. (2025). Two-step multilevel latent class analysis in the presence of measurement non-equivalence. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–10.
MacLahlan, & Peel. (2000). *Finite mixture models.* John Wiley & Sons.
Magidson. (1981). Qualitative variance, entropy, and correlation ratios for nominal dependent variables. *Social Science Research*, *10*(2), 177–194.
Malsiner-Walli, Grün, & Frühwirth-Schnatter. (2025). Without pain–clustering categorical data using a Bayesian mixture of finite mixtures of latent class analysis models. *Advances in Data Analysis and Classification*, 1–36.
Masyn. (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. (1979). *Analysis of qualitative data: New developments.* Academic Press.
Millsap. (2011). *Statistical approaches to measurement invariance.* New York: Routledge.
Morgan. (2015). Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 76–86.
Morin, Legault, Laliberté, Bakk, Giguère, de la Sablonnière, & Lacourse. (2023). StepMix: A Python package for pseudo-likelihood estimation of generalized mixture models with external variables. *arXiv preprint arXiv:2304.03853*.
Muthén, & Muthén. (2017). Mplus. In *Handbook of item response theory* (pp. 507–518). Chapman

and Hall/CRC.
National Endowment for the Arts. (2015). *Local Area Arts Participation Study 1992.* Inter-university Consortium for Political and Social Research [distributor]. Retrieved from https://doi.org/10.3886/ICPSR35593.v1
Nylund, Asparouhov, & Muthén. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.
Oser. (2022). Protest as one political act in individuals' participation repertoires: Latent class analysis and political participant types. *American Behavioral Scientist*, *66*(4), 510–532.
Oser, Di Mari, & Bakk. (2023). Data preparation for citizenship norm analysis, International Association for the Evaluation of Educational Achievement (IEA) 1999-2009-2016. *Open Science Framework*, *10*. Retrieved from https://doi.org/10.17605/OSF.IO/AKS42
Oser, & Hooghe. (2013). The evolution of citizenship norms among Scandinavian adolescents, 1999–2009. *Scandinavian Political Studies*, *36*(4), 320–346.
Oser, Hooghe, Bakk, & Di Mari. (2023). Changing citizenship norms among adolescents, 1999-2009-2016: A two-step latent class approach with measurement equivalence testing. *Quality & Quantity*, *57*, 4915–4933.
Peracchi. (2001). *Econometrics.* John Wiley & Sons.
Prior. (2010). You've either got it or you don't? the stability of political interest over the life cycle. *The Journal of Politics*, *72*(3), 747–766.
Qiu. (2022). A tutorial on Bayesian latent class analysis using jags. *Journal of Behavioral Data Science*, *2*(2), 127–155.
R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual].
Rabe-Hesketh, & Skrondal. (2022). *Multilevel and longitudinal modeling using Stata* (4th ed.). College Station, TX: Stata Press.
Raftery, & Lewis. (1992). How many iterations in the Gibbs sampler. *Bayesian Statistics*, *4*(2), 763–773.
Rosseel, & Loh. (2024). A structural after measurement approach to structural equation modeling. *Psychological Methods*, *29*, 561–588.
Schulz, Ainley, Fraillon, Losito, Agrusti, & Friedman. (2018). *Becoming citizens in a changing world: IEA International Civic and Citizenship Education Study 2016 international report.* Springer Nature.
Throsby. (1994). A work-preference model of artist behaviour. In *Cultural economics and cultural policies* (pp. 69–80). Springer.
Van Deth, Abendschön, & Vollmar. (2011). Children and politics: An empirical reassessment of early political socialization. *Political Psychology*, *32*(1), 147–174.
Van Rees, Vermunt, & Verboord. (1999). Cultural classifications under discussion latent class analysis of highbrow and lowbrow reading. *Poetics*, *26*(5-6), 349–365.
Vermunt. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213–239.
Vermunt. (2005). Mixed-effects logistic regression models for indirectly observed discrete outcome variables. *Multivariate Behavioral Research*, *40*(3), 281–301.
Vermunt. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469.
Vermunt, & Magidson. (2021a). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 356–364.
Vermunt, & Magidson. (2021b). LG-syntax user's guide: Manual for Latent GOLD syntax module version 6.0. *Arlington, MA: Statistical Innovations*.
Vidotto, Vermunt, & Van Deun. (2018). Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, *43*(5), 511–539.
White, & Murphy. (2014). Bayeslca: An R package for Bayesian latent class analysis. *Journal of Statistical Software*, *61*, 1–28.

Wyse, & Friel. (2012). Block clustering with collapsed latent block models. *Statistics and Computing*, *22*, 415–428.

Xue, & Bandeen-Roche. (2002). Combining complete multivariate outcomes with incomplete covariate information: A latent class approach. *Biometrics*, *58*, 110–120.

Zhang, Kuha, & Steele. (2024). Modelling correlation matrices in multivariate data, with application to reciprocity and complementarity of child-parent exchanges of support. *The Annals of Applied Statistics*, *18*(4), 3024–3049.