Two-stage multilevel latent class analysis with co-variates in the presence of direct effects

Zsuzsa Bakk
Department of Methodology and Statistics, Leiden University, The Netherlands

Roberto Di Mari
Department of Economics and Business, University of Catania, Italy

Jennifer Oser
Department of Politics and Government, Ben-Gurion University, Israel

Jouni Kuha
Department of Statistics, London School of Economics, UK

Abstract

In this article we present a two-stage estimation approach applied to multilevel latent class analysis (LCA) with co-variates. We separate the estimation of the measurement and structural model. This makes the extension of the structural model computationally efficient. We investigate the robustness against misspecifications of the proposed two-stage and the classical one-stage approach for models where a direct effect exists between indicators of the LC model and covariate, and the direct effect is ignored.

Keywords: multilevel latent class analysis; covariates; two-stage estimation; direct effects; robustness, misspecifications

z.bakk@fsw.leidenuniv.nl

Introduction

Latent class (LC) analysis is an approach used to create a clustering of a set of observed variables, based on an underlying unknown classification. For example based on indicators such as intensity and type of internet use Hsieh & Yang (2011) used LC analysis to identify distinctive clusters of internet usage segments in Taiwan, such as business, amusement, entertainment and online shopping, and leisure. In *multilevel* LC analysis the respondents are assumed to belong to higher level groups, such as students nested in schools, or entrepreneurs in countries. Using *multilevel* LCA the higher level dependency is modeled by assuming that respondents nested in the same higher level unit have more similar answers to each other than respondents coming from different units. For example Hsieh & Yang (2011) found that the internet user profiles in Taiwan can be clustered into three segments: Southern Taiwan, Northern Taiwan, and metropolitan.

Multilevel LCA is becoming increasingly popular in various fields. For instance, in educational research, to model students' learning profiles in different school types (Fagginger Auer et al., 2016), or to model academic profiles (Lanza et al., 2010; Mutz et al., 2013) or to cluster psychology students in different attitude types towards learning statistics and at the same time obtaining university segments based on the incidences of the different student attitude types (Mutz & Daniel, 2013); in economics, to model asset ownership types of the elderly across Europe (Paccagnella & Varriale, 2013); or epidemiology and health studies to model substance abuse profiles nested in different communities (Rindskopf, 2006; Horn et al., 2008; Zhang et al., 2012; Tomczyk et al., 2015), to mention a few. Some further examples from political science include the modeling of heterogenity of what Europeans think is the cause of poverty (Da Costa & Dias, 2015), or changes in social capital over time (Morselli & Glaeser, 2018) or a typology of trust orientation towards European institutions (Ruelens & Nicaise, 2020). In most applications the interest lies at the lower level clustering, and the difference in the distribution of the lower level classes at the higher level unit.

In LCA creating a clustering is usually only the first step for applied researchers. The research interest often lies in explaining the clustering by co-variates. Examples include relating heavy alcohol usage profiles to age, gender, education and religion (Rindskopf, 2006), or teen dating violence in China to demographic characteristics (Cheng et al., 2020).

While in single level LCA different approaches are available for relating LC membership to external variables, in multilevel settings only two classical approaches are used, both known to be suboptimal, namely the one-step and classical three-step approaches. Using the one-step approach the full LC model including co-variates is estimated simultaneously (for example, Mutz & Daniel (2013)). Using the alternative three-step approach, after estimating the measurement model in step 1, respondents are assigned to latent classes in step 2, and this posterior assigned class membership is related to the predictors of interest through a multinomial logit regression in the third step (for example Tomczyk et al. (2015)). However in the second step a classification error is introduced, that if not corrected for induces systematic bias in the step 3 model.

To correct for the bias in the step 3 model of the three-step approach, in recent years two bias-adjusted three-step approaches were developed for single level LC models - namely the ML and BCH approaches (Bakk et al., 2013; Vermunt, 2010). The bias-adjusted three-step approaches correct the bias in step 3 by explicitly modeling the classification error introduced in the previous step. An alternative stepwise estimator, the two-step approach (Bakk & Kuha, 2018) after estimating the measurement model in step one, directly conditions on the step one parameter estimates in the second step when estimating the structural model, in this way avoiding the problem of the classification error.

The general recommendation in single level LCA is to use the two-step or bias-adjusted three-step approaches to relate the LC measurement model to external variables of interest (Asparouhov & Muthén, 2014), with the understanding that the two-step approach is the most flexible to extend to more complex models (Bakk & Kuha, 2018; Di Mari & Bakk, 2018). The main reason for using these stepwise estimators instead of the one-step approach is that misspecifications in the structural model can influence the definition of the measurement model using the one-step approach. For example direct effects between the co-variate and some indicators measuring the LC variable can distort the parameters of interest, or perhaps can have an even more worrisome impact on latent class enumeration - extracting more classes then necessary to model the direct effects (also known as differential item functioning, DIF) (Cole et al., 2019; Masyn, 2017). Because they separate measurement and structural model the bias adjusted stepwise approaches are known to be more robust to misspecifications.

In the current paper we introduce the two-stage approach to *multilevel* LC modeling as an alternative to the one-step and classical three-step approaches, since both are known to be suboptimal in single level LC models. The proposed two-stage estimator separates each step of the model building, namely first the lower level LC model is built. Next while keeping the lower-level measurement model fixed the higher level mixing proportions are selected. Finally conditioning on the fixed parameter estimates of the two-level measurement model the structural model is estimated. We investigate the robustness of the one-step and two-stage approaches towards misspecifications of the co-variate effect. We focus on one of the most common misspecifications, that is ignoring direct effect(s) between the covariate and indicators. Via a simulation study we investigate the performance of the one and two-stage approaches with regard to bias and MSE when modeling and ignoring the direct effects. We also investigate Type 1 error rate for models that misspecified the relationship between the external variable of interest and the measurement model.

First we present the measurement model of the multilevel latent class model, and the inclusion of covariates using the one and two-stage approaches. We discuss the inclusion of direct effects between the covariates and indicators for both modeling approaches, and following in a simulation study we investigate the impact of misspecification of direct effects on the two modeling approaches under different levels of violations of the assumption of local independence. We apply both approaches to a real data setting, and we conclude.

The multilevel latent class model

Consider the vector of responses $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})$, where Y_{ijk} denotes the response of individual i in group j on the k-th categorical indicator variable, with $1 \leq k \leq K$ and $1 \leq j \leq J$, where K denotes the number of categorical indicators and J the number of level 2 units (groups). In addition, we let n_j denote the number of level 1 units within the j-th level 2 unit, with $1 \leq j \leq J$. For simplicity of exposition, we focus on dichotomous indicators.

LC analysis assumes that respondents belong to one of the T categories (“latent classes”) of an underlying categorical latent variable X which affects the responses (McCutcheon, 1987; Goodman, 1974; Hagenaars, 1990). The model for $\mathbf{Y}_{ij}$ can then be specified as

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

where the weight $P(X_{ij} = t)$ is the probability of person i in group j to belong to latent class t.

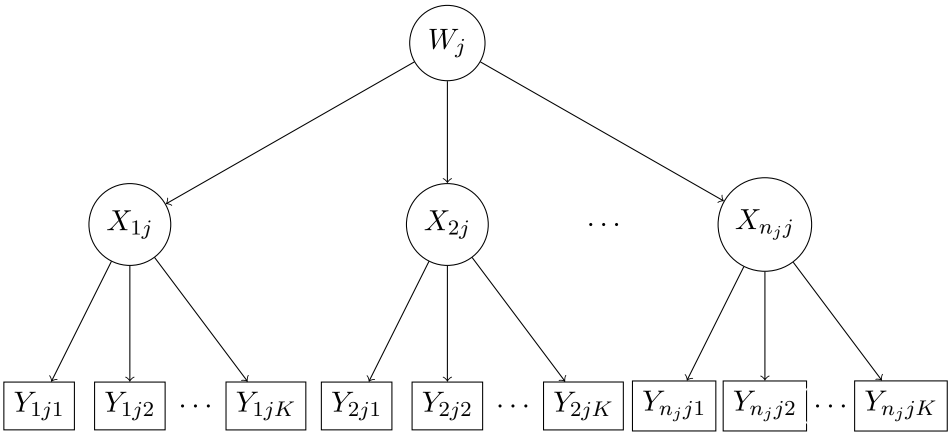


Figure 1. The *multilevel* latent class model.

The term $P(\mathbf{Y}_{ij}|X = t)$ is the class-specific probability of observing a pattern of responses given that a person belongs to class t. Furthermore we make the "local independence" assumption that the K indicator variables are independent within the latent classes, leading to

$$P(\mathbf{Y}_{ij}) = \sum_{t=1}^{T} P(X_{ij} = t) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t). \tag{2}$$

Note that the general definition in Equation (1) applies to both the standard and *multilevel* LC model. To be able to distinguish the simple and *multilevel* LC model we can define the model in terms of logit equations. In the simple LC model

$$P(X_{ij} = t) = \frac{\exp(\gamma_t)}{1 + \sum_{t=2}^{T} \exp(\gamma_t)}, \tag{3}$$

for $1 < t \leq T$ - where we have taken the first class as reference - and

$$P(Y_{ijk} = 1|X_{ij} = t) = \frac{\exp(\beta_t^k)}{1 + \exp(\beta_t^k)}. \tag{4}$$

In the simple LC model the parameters γ and β do not have the subscript j, thus assuming the clustering is independent of the higher level groups.

Extending the simple LC model to account for the multilevel data structure is possible by allowing the parametrizations (3) and (4) to take the grouping (level 2 units) into account by means of group-specific random coefficients. As such, the *multilevel* LC model can be seen as a random coefficients logistic regression model (see, for instance Agresti et al., 2000) for an unobserved dependent variable, which has several observed indicators (Vermunt, 2003). Therefore, the parametrization of *multilevel* LC models can follow either the *parametric* approach or the *nonparametric* approach (see also Finch & French, 2014). In the parametric approach, group-specific effects are assumed to arise from a certain continuous distribution, typically Gaussian. In the nonparametric approach (Aitkin, 1999; Laird, 1978), instead of a continuous distribution we assume a multinomial distribution. Let W_j denote the value of group j on the latent class variable defining the mixing distribution with M mass points each with probability $P(W_j = m) = \pi_m$. In the nonparametric approach the model for the (individual) latent class probabilities is specified (Vermunt, 2003):

$$P(X_{ij} = t|W_j = m) = \frac{\exp(\gamma_{tm})}{1 + \sum_{s=2}^{T} \exp(\gamma_{sm})}. \tag{5}$$

Also the mixing probabilities $P(W_j = m)$ can be parametrized by means of logistic regressions as follows

$$P(W_j = m) = \frac{\exp(\delta_{0m})}{1 + \sum_{l=2}^{M} \delta_{0l}}, \tag{6}$$

where parameters for $m = 1$ are set to zero for identification and the related class is set as reference. This is the most commonly used specification in applied research due to its simplicity.

Following the same logic, the conditional response probabilities of Equation (4) become

$$P(Y_{ijk} = 1|X_{ij} = t, W_j = m) = \frac{\exp(\beta^k_{tm})}{1 + \exp(\beta^k_{tm})}, \tag{7}$$

for $k = 1, \ldots, K$, $t = 1, \ldots, T$ and $m = 1, \ldots, M$. This is the most general formulation that is equal to an unrestricted multi-group LC model. In most applications, however, a more restricted version is used (Vermunt, 2003; Lukociene et al., 2010) that assumes that item-conditional probabilities (see Equation (7)) do not depend on the level 2 units – as in Equation (4). This is the restriction we will apply also in the current paper (see Figure 1), leading to the following specification for $\mathbf{Y}_{ij}$:

$$P(\mathbf{Y}_{ij}) = \sum_{m=1}^{M} P(W_j = m) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t). \tag{8}$$

Under the parametrizations (3), (7) and (6), given a sample of observations from J groups - each with n_j individual units, for $j = 1, \ldots, J$, with $N = \sum_{j=1}^{J} n_j$ - the log likelihood function for model (8) can be written as:

$$\log L(\theta) = \sum_{j=1}^{J} \log P(\mathbf{Y}_{ij}), \tag{9}$$

which we maximize in order to find the vector of model parameters θ. The numbers of classes M and T are selected by comparing the goodness of fit of models with different values of M and T using information criteria like AIC and BIC.

All the parts of the *multilevel* LC model can be estimated simultaneously. However the choice of the number of latent classes on level 1 and level 2 is not so obvious. A generally used recommendation is to use a stepwise procedure for model selection (Lukociene et al., 2010), by first fitting a single-level LC model at the level 1 - defined in Equations 3 and 4. Once the correct number of classes at the lower level is selected, this number is held fixed and the number of classes is estimated at the higher level. A general recommendation is that, once the higher level classes are selected, these are kept fixed, and model selection is re-iterated at the lower level one more time before adding covariates. In the stage of adding covariates the number of classes should be fixed, also to be in line with general recommendations for LCA with covariates (Masyn, 2017). Readers interested in model selection for *multilevel* LCA model can also consult Yu & Park (2014).

Extending the *multilevel* LC model to include covariates

Classical approaches

Level 1 and level 2 covariates can be included to predict class membership. Denoting one level 2 covariate by Z_{1j} and a level 1 covariate by Z_{2ij} the multinomial logistic regression for X_{ij} with a random intercept can be written as:

$$P(X_{ij} = t|W_j = m, Z_{1j}, Z_{2ij}) = \frac{\exp(\gamma_{0tm} + \gamma_{1t}Z_{1j} + \gamma_{2t}Z_{2ij})}{1 + \sum_{s=2}^{T} \exp(\gamma_{0sm} + \gamma_{1s}Z_{1j} + \gamma_{2s}Z_{2ij})}. \tag{10}$$

A random slope for the level 1 covariate can be obtained by replacing γ_{2t} by γ_{2tm}.

Level 2 covariates can be used also to predict group class membership. To do so, the multinomial logistic regression for $P(W_j = m)$ can be modified as follows

$$P(W_j = m|Z_{1j}) = \frac{\exp(\delta_{0m} + \delta_{1m}Z_{1j})}{1 + \sum_{l=2}^{M} \exp(\delta_{0l} + \delta_{1l}Z_{1j})}. \tag{11}$$

Under the parametrizations (10) and (11) that now include covariates, the model for $\mathbf{Y}_{ij}|\mathbf{Z}_j$, where $\mathbf{Z}_j = (Z_{1j}, Z_{2ij})'$, can be specified as

$$P(\mathbf{Y}_{ij}|\mathbf{Z}_j) = \sum_{m=1}^{M} P(W_j = m|Z_{1j}) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, Z_{1j}, Z_{2ij}) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t), \tag{12}$$

where we have further assumed that the observed indicators are conditionally independent from the covariates given both level 1 and level 2 class memberships.

Using the one-step approach the full model needs to be re-estimated every time a new covariate is added keeping the number of lower and higher level classes fixed. Given the complexity of such multilevel models, 1) estimating the full model multiple times can be time consuming, and 2) misspecifications in a part of the model may destabilize also parameters in other parts of the model.

Two-stage estimation of multilevel LC models

An alternative option that would fit the logic of the stepwise modeling procedure is to apply a two-stage estimation approach by extending the two-step approach proposed for simple LC models by Bakk & Kuha (2018) and applied to latent Markov models by Di Mari & Bakk (2018). We apply the two-step logic in the multilevel context using a stage-wise approach. Namely, first the lower level LC model is estimated (step 1 see Figure 2). Once the number of lower level classes are selected the higher level LC model is estimated keeping the measurement model fixed at the estimates from the previous step 1 (Step 2a see Figure 3). In this way only the mixing proportions (at both levels) need to be re-estimated, keeping the $P(\mathbf{Y}_{ij}|X_{ij})$ fixed at the values estimated in step 1. Similarly to the simultaneous estimation, once the higher level LC model is selected, keeping this part fixed, the model for $P(\mathbf{Y}_{ij}|X_{ij})$ can be re-estimated to adjust for possible missspecifications due to grouping at level 2 (Step 2b see Figure 4). Finally, the covariates can be added to

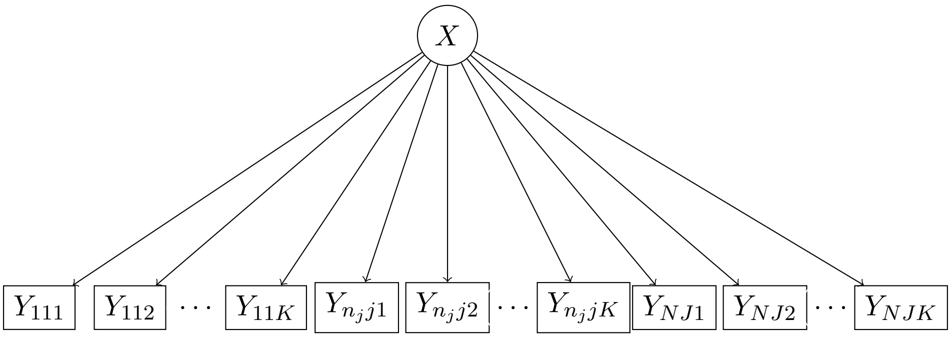


Figure 2. Stage 1 Step 1: simple latent class model on the pooled observations - multilevel structure of the data is not taken into account. This step is equivalent to simple LCA on the pooled observations.

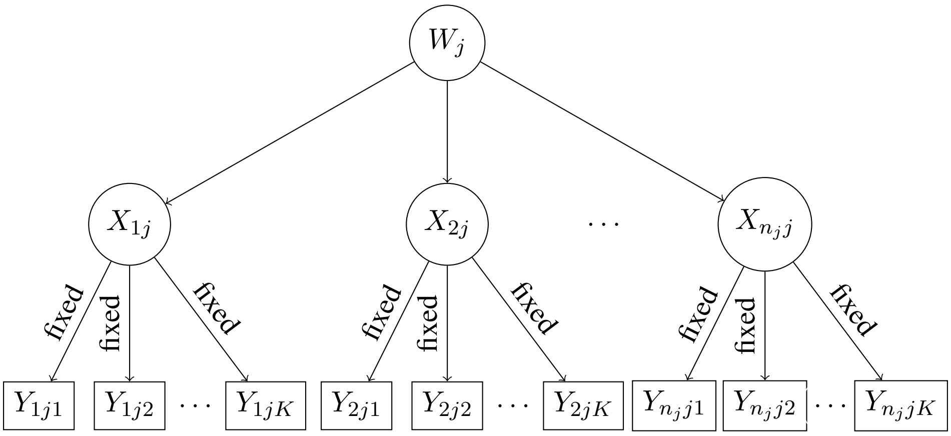


Figure 3. Stage 1 Step 2.a: multilevel latent class model with measurement model kept fixed at Step 1's values.

the model keeping the measurement model fixed (Step 3 see Figure 5) . In the next section we describe in detail each step of the proposed estimator. We shall distinguish between the steps without covariates (steps 1, 2a and 2b) - which we refer to as stage 1 of *multilevel* LC model building - and the step(s) with covariates (step 3) - which we refer to as stage 2.

The steps of the two-stage estimation for *multilevel* LC model with covariates

Stage 1: Unconditional LC model building

Step 1: Simple LC model. In this step (Figure 2) a simple LC model is estimated on the pooled data T_{max} times, where T_{max} is a pre-specified maximum number of latent classes for X. We let $\theta_1^{(T)} = (\beta^1_{2_1}, \ldots, \beta^1_{T_1}, \ldots, \beta^1_{T_J}, \ldots, \beta^K_{2_1}, \ldots, \beta^K_{T_1}, \ldots, \beta^K_{T_J})'$ for each choice of $T = 1, \ldots, T_{\text{max}}$. Under the parametrizations (3) and (4), and a sample of N observations - where $N = \sum_{j=1}^{J} n_j$ - the log likelihood function of the first step model can be specified as follows

$$\log L(\theta_1) = \sum_{i=1}^{N} \log P(\mathbf{Y}_{ij}), \tag{13}$$

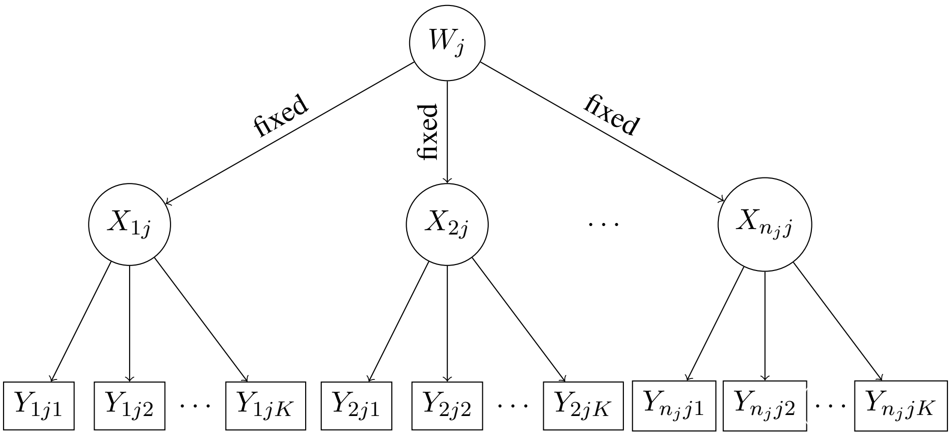


Figure 4. Stage 1 Step 2.b: measurement model is updated to account for possible interaction effects with high level parameters.

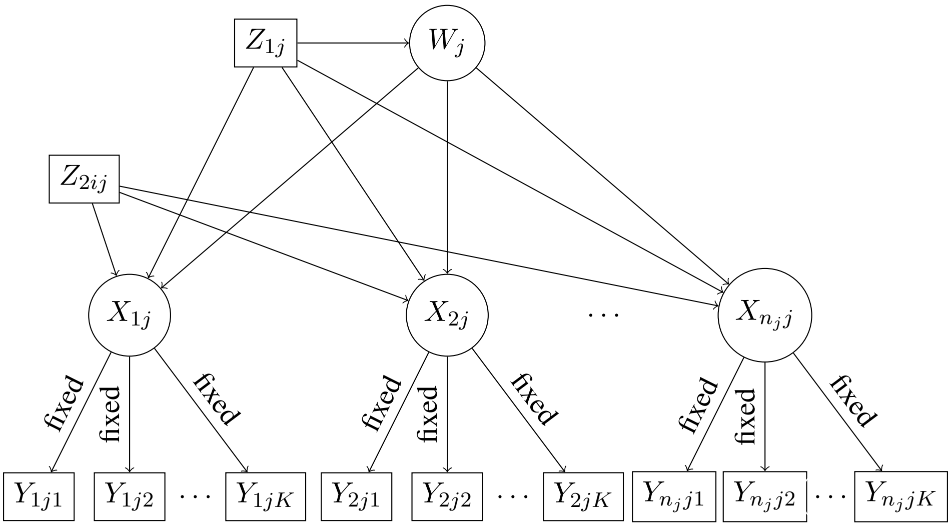


Figure 5. Stage 2 Step 3: covariates Z_{1j} and Z_{2ij} are loaded on X_ij and W_j, keeping measurement model parameters fixed. Note that this step can be carried out either simultaneously or in two separate sub-steps: 3.a) covariates loaded only in the equations for X_{ij}, for all j; 3.b) keeping the parameters estimated in 3.a fixed, covariates are loaded on W_j .

which we maximize in order to find the ML estimates of the LC model parameters, which we call $\widehat{\theta}_1^{(T)} = (\widehat{\beta}^1_{2_1}, \ldots, \widehat{\beta}^1_{T_1}, \ldots, \widehat{\beta}^1_{T_J}, \ldots, \widehat{\beta}^K_{2_1}, \ldots, \widehat{\beta}^K_{T_1}, \ldots, \widehat{\beta}^K_{T_J})'$. Then, the optimal number of classes T^* is selected such that $T^* = \min_{T=1,\ldots,T_{\max}} I(T)$ - where $I(T)$ is some information criterion like AIC or BIC - along with the corresponding ML estimates $\widehat{\theta}_1$ from which we have suppressed the superscript $^{(T)}$ for simplicity of notation.

Step 2.a: *multilevel* LC model. In this step (Figure 3), the group level measurement model parameters of the *multilevel* LC model are estimated keeping measurement model parameters at the lower level fixed at $\widehat{\theta}_1$. In the same way as for Step 1, this step has to be carried out $M_{\max}$ times, where $M_{\max}$ is a pre-specified by the user number of latent classes for W. We let

$\theta_2^{(M)} = (\delta_2, \ldots, \delta_M, \ldots, \gamma_{021}, \ldots, \gamma_{0T1}, \ldots, \gamma_{0TM})'$ for each choice of $M = 1, \ldots, M_{\text{max}}$. Under the parametrizations (4), (5) and (6), and a sample of J groups - each with n_j individual units, for $j = 1, \ldots, J$ - the log likelihood function of the step 2.a model can be written as follows

$$\log L(\theta_2 | \theta_1 = \widehat{\theta}_1) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j), \tag{14}$$

where $|\theta_1 = \widehat{\theta}_1$ indicates that the measurement model parameters are kept fixed at $\widehat{\theta}_1$. The function (15) is maximized with respect to the unknown θ_2 to find ML estimates $\widehat{\theta}_2$.

Step 2.b: Re-update the measurement model. In this step (Figure 4) the level 1 measurement model is re-estimated, keeping fixed the level 2 model parameters. This is done in order to re-adjust the lower- level measurement model if necessary based on the selected number of higher level classes. Note that in principle, as for step 1, this step can be carried out T_{max} times - i.e. freeing also the parameters on X's equations and re-estimating the optimal number of low level classes. Such a *full* step maybe unnecessary in most situations (see Lukociene et al. (2010)).

Given a sample of J groups - each with n_j individual units, for $j = 1, \ldots, J$ - and vector of estimates $\widehat{\theta}_2$ from the previous step, under the parametrizations (4), (5) and (6), the log likelihood function of the step 2.b model can be written as

$$\log L(\theta_1 | \theta_2 = \widehat{\theta}_2) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j). \tag{15}$$

This specification, with respect to that of Step 1, now takes the multilevel structure of the data into account.

Stage 2: including covariates

Step 3: including predictors for class memberships. As the next step the covariates can be added to the model (Figure 5). A decision needs to be taken whether a stepwise approach is preferred (adding first lower level covariates, and after fixing those adding at the higher level) or all covariates can be added in a single step. The benefit of the first option can be robustness, however no simulation or theoretical results are available and this still needs further research. For sake of conciseness, we will present the simultaneous step; its split counterpart can be derived analogously.

Let us define $\theta_3 = (\gamma_{12}, \ldots, \gamma_{1T}, \gamma_{22}, \ldots, \gamma_{2T})'$. With the parametrizations specified in Equations (4), 10 and (11), the model log-likelihood can be written as follows

$$\log L(\theta_2, \theta_3 | \theta_1 = \widehat{\theta}_1) = \sum_{j=1}^{J} \log P(\mathbf{Y}_j | \mathbf{Z}_j), \tag{16}$$

which we maximize with respect to θ_2 and θ_3 keeping θ_1 fixed at its step 2.b values $\widehat{\theta}_1$.

Modeling direct effects between covariates and indicators of the LC model

While the model defined in Equation (12) assumes conditional independence between the indicators and the covariate given the latent class variable, in some cases this assumption can be violated. Such violation is also known as differential item functioning (DIF). Such violation exists for example when an item has a different difficulty for boys and girls in an educational tests (lower level Z) or an item has different difficulty in different countries (higher level Z). The LC model can be extended to relax the conditional independence assumption, by allowing a direct effect. Keeping the assumption that the measurement model does not depend on W we modify Equation 2 as:

$$P(\mathbf{Y}_{ij} = 1) = \sum_{t=1}^{T} P(X_{ij} = t) \prod_{k=1}^{K} P(Y_{ijk}|X_{ij} = t, Z_{1j}, Z_{2ij}). \tag{17}$$

The model defined in equation 17 can be expressed in terms of a logit equation:

$$P(Y_{ijk}|X_{ij} = t, Z_{1j}, Z_{2ij}) = \frac{\exp(\beta_t^k + \beta_{1tk} Z_{1j} + \beta_{2tk} Z_{2ij})}{1 + \exp(\beta_t^k + \beta_{1tk} Z_{1j} + \beta_{2tk} Z_{2ij})}. \tag{18}$$

Equation 18 defines the most general form to allow for direct effect on the indicators from covariates at the lower and/ or higher level.

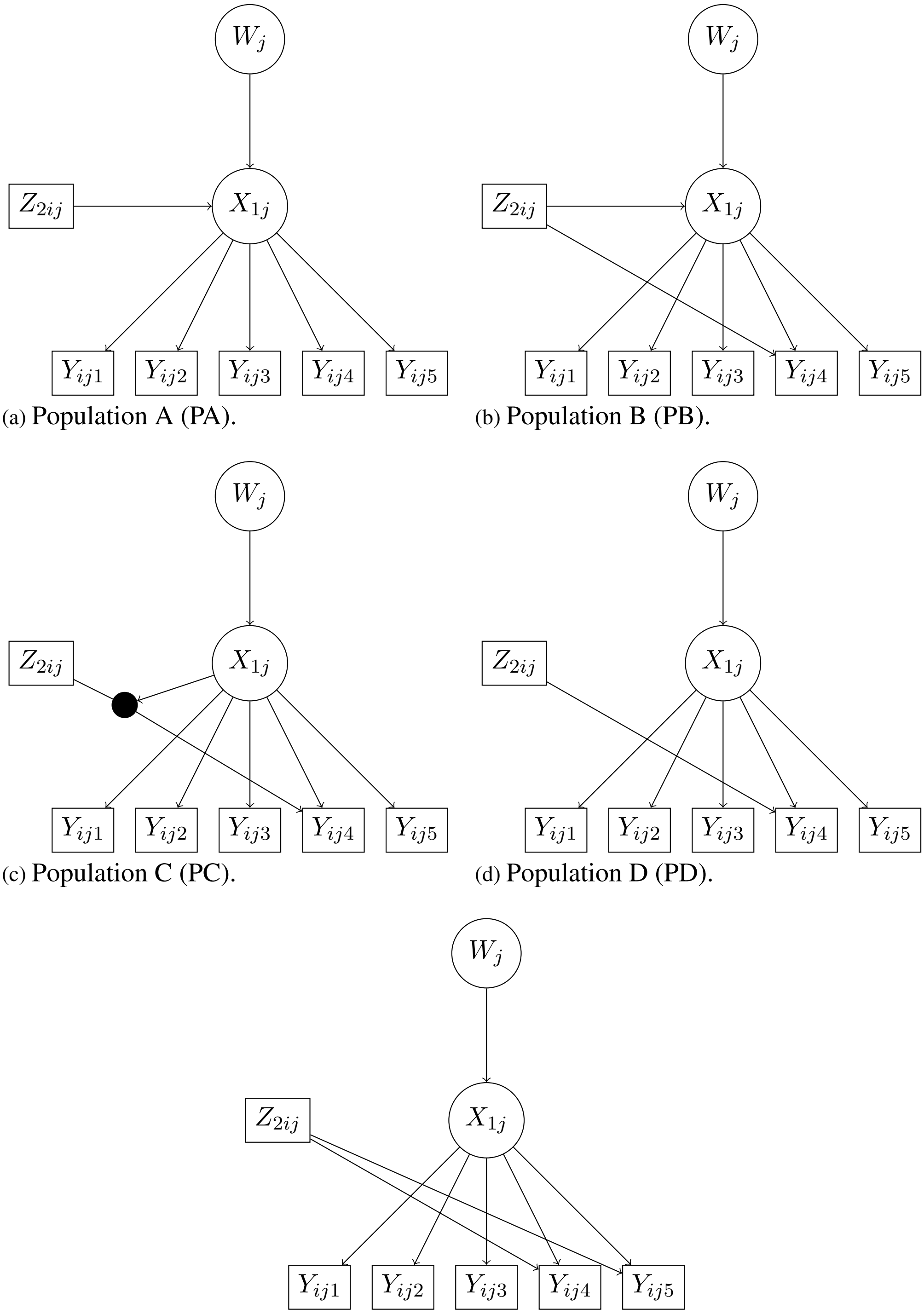
Using the one-step approach the full LC model is estimated allowing for all necessary direct effects. Using the two-stage approach on the other hand the measurement model is kept fixed at the estimates from Stage 1 step 2b for the indicators for which no DIF is assumed, and the conditional item probabilities are re-estimated using Equation 18 for the indicators for which the assumption of DIF is being relaxed. In this way the two-stage approach is more parsimonious. Using the classical or even the bias-adjusted three-step approaches the modeling of DIF is not possible.

While modeling direct effects with both one and two-stage approaches is possible this is often not done in practice. The reasons for these are diverse: most importantly increased model complexity makes interpretation more difficult. Furthermore detecting direct effects is difficult. The literature recommends using overall fit statistics or residual statistics (Oberski et al., 2013), but no clear consensus exists about the power of detecting such effects for *multilevel* LC models (Nagelkerke et al., 2015).

In the current paper we focus on understanding the effect on parameter bias of the parameters of interest ($X|Z$) if direct effects are ignored.

Simulation study

We carry out a simulation study to investigate the performance of the proposed two-stage estimator as compared to the simultaneous estimator with regard to bias and efficiency. Next to the situation where all model assumptions are met we also investigate the impact of ignoring direct effect(s) in *multilevel* LCA. For this purpose we generated data from 5 population models with different types of direct effects. We followed the setup by Nylund-Gibson & Masyn (2016), who investigated the impact of DE misspecification on class enumeration for single level models. We go a step further and investigate the impact on parameter bias in multilevel setting. Two of the five settings are population models where only indirect effect between X, Z exist via a direct effect of Z on indicator(s) Y. While this situation can be a common population model, it is hardly used in data analysis, as most models include the direct X, Z association. As such investigating how modeling the X, Z association while ignoring the true Y, Z association shows Type 1 error rates in such complex settings.



(a) Population A (PA).

(b) Population B (PB).

(c) Population C (PC).

(d) Population D (PD).

(e) Population E (PE).

Figure 6. Population models.

Population models

We generate data from 5 models with DE for 2 separate Z variables, namely at the lower and higher level, so that we have 10 data generating models. The 5 models are presented in Figure

6 for the Z at lower level. The same effect size measures are used for Z at the higher level as well. All the population models were specified to be measured by 5 binary indicator variables, measuring 2 classes at the lower level with all indicators having a high probability of a positive answer ($P(Y|X) = 0.80$)) in one class, and low in the other class ($P(Y|X) = 0.20$)). At the higher level we have $W = 2$ classes. We have two scenarios for class sizes, with equal and unequal class sizes at both the lower and higher level. In the equal class sizes scenario the class sizes were set at $\pi_1 = \pi_2 = 0.5$ at both the lower and higher level, and in the unequal class sizes scenario to $\pi_1 = 0.8, \pi_2 = 0.2$. We generated data from a population with 20 groups at the higher level, and with sample size n_j 500 and 1000 at the lower level.

In model PA, PB and PC, β_{1t} for the effect of Z_{2ij} on Y_4 was set to 1 in the models with a lower level covariate. In models with a higher level covariate $\beta_{2t}Z_{1j}$ was set to 1 as well. In model PB, PD and PE the size of the direct effect of Z on Y_4, was set to 1 for both the lower ($\beta_{1t4}Z_{2ij}$) and higher level ($\beta_{2t4}Z_{1j}$) setting. In model PC the direct effect in class one was set to 0.5, and in class two to 1.5.

Analysis models

In all instances we analyze the data using the PA model (which is the most common model used in practice) and the true data generating model, using the one and two-stage approaches. R is used for data manipulation and Latent GOLD for the estimation of the LC models.

Results

In Table 1 we show the simulation results for the data generating models PA, PB and PC for the data generated with equal class sizes at both levels, and group level Z. We zoom in on the main effect of Z on X, and show for each data generating condition the results with the PA model (ignoring DE between Z and Y), and the true data generating model. The results under the PA data generating model show that when all model assumptions are met both the one and two-stage estimators are unbiased, with very similar coverage (somewhat above the nominal 95 % rate), MSE and SE/SD.

For the data generating models PB and PC analyzing the data ignoring the direct effect introduces a bias in the parameter of interest with both the one and two-stage approaches (the effect size of the bias is marginally smaller with the two-stage approach). Interestingly enough the coverage even with the PA model is above the nominal level, possibly due to a larger SE estimate.

Table 2 presents some results for data generated under models PD and PE, that have a direct effect between Z, Y, but no relationship between Z, X. Analyzing these populations with model PA, which assumes no direct effect between Z and any of the indicators, only indirect effect via X sheds light on the Type 1 error using the wrong PA model. As the last column of Table 2 shows this Type 1 error rate is above 5% in all situations, and with the number of direct effects increasing it can go as high as 56% in the conditions presented. The Type 1 error rate is also higher under both the PE and PD data generating model for the larger sample size. Using the correct data generating model to analyze the data with both one and two-stage approaches the bias is negligible in the parameters of interest, and the coverage rate above the nominal rate. This is also the case for the models PA to PC.

Table 3 shows the results for data generated from models PA to PC for uneven class sizes with Z at higher level. The results are similar to the previous condition with regard to bias, cover-

Data analysis model/	One-step				Two-stage			
sample size				Data generating model: PA				
PA model	Bias	Coverage	MSE	SE/SD	Bias	Coverage	MSE	SE/SD
500	0.01	0.96	0.05	0.94	0.01	0.96	0.05	0.94
1000	0.00	0.97	0.02	1.00	0.00	0.97	0.02	1.00
				Data generating model: PB				
PA model								
500	-0.17	0.99	0.08	0.93	-0.15	0.99	0.07	0.94
1000	-0.16	1.00	0.05	0.96	-0.14	1.00	0.04	0.97
PB model								
500	-0.01	0.98	0.05	0.95	-0.01	0.98	0.05	0.95
1000	0.00	0.97	0.02	0.99	0.01	0.97	0.02	0.98
				Data generating model: PC				
PA model								
500	-0.11	1.00	0.06	0.95	-0.09	1.00	0.06	0.96
1000	-0.11	0.99	0.04	0.99	-0.10	0.99	0.03	0.99
PC model								
500	0.01	0.98	0.05	0.98	0.01	0.98	0.05	0.98
1000	0.01	0.97	0.02	0.99	0.01	0.97	0.02	0.99

Table 1

Parameter bias, Coverage, MSE and SE/SD for $\gamma_1 = 1$ measuring the direct effect of $X|Z$ for data generated from models PA, PB, and PC, analyzed with model PA for even class sizes, group level Z.

Model	Sample size	Bias	Coverage	MSE	Type 1E
			Data generating model PD		
			PD model		PA model
One-step	500	-0.01	0.97	0.06	0.11
	1000	0.00	0.98	0.03	0.16
Two-stage	500	0.00	0.97	0.06	0.11
	1000	0.00	0.97	0.03	0.15
			Data generating model PE		
		PE model	Average effects		PA model
One-step	500	-0.02	0.99	0.07	0.29
	1000	-0.02	0.98	0.04	0.56
Two-stage	500	-0.02	0.99	0.07	0.29
	1000	-0.01	0.97	0.04	0.56

Table 2

(Average) Bias, Coverage and MSE of the Direct effect(s) of Z on Y for models PD and PE, and Type 1 error rate for $P(X|Z)$ for analyzing data generated from Models PD and PE with model PA even class sizes, group level Z.

Data analysis model/	One-step				Two-stage			
sample size				Data generating model: PA				
PA model	Bias	Coverage	MSE	SE/SD	Bias	Coverage	MSE	SE/SD
500	-0.01	0.96	0.03	1.00	-0.00	0.96	0.03	0.94
1000	0.00	0.97	0.01	1.00	0.01	0.97	0.01	1.00
				Data generating model: PB				
PA model								
500	0.11	0.86	0.04	1.00	0.13	0.83	0.04	0.94
1000	0.10	0.82	0.02	0.92	0.12	0.77	0.03	0.92
PB model								
500	-0.00	0.97	0.01	1.00	0.01	0.97	0.03	1.00
1000	-0.01	0.97	0.01	0.99	-0.00	0.97	0.01	1.00
				Data generating model: PC				
PA model								
500	0.10	0.87	0.04	1.00	0.13	0.83	0.04	0.94
1000	0.09	0.85	0.02	0.92	0.11	0.78	0.03	0.92
PC model								
500	-0.00	0.97	0.03	1.00	0.01	0.96	0.03	1.00
1000	-0.01	0.98	0.01	1.00	-0.01	0.98	0.01	1.00

Table 3

Parameter bias, Coverage, MSE and SE/SD for $\gamma_1 = 1$ measuring the direct effect of $X|Z$ for data generated from models PA, PB, and PC, analyzed with model PA and the data generating model for the two levels of sample size conditions. Uneven group-level class sizes group level Z.

Model	Sample size	Bias	Coverage	MSE	Type 1 E
			Data generating model PD		
			PD model		PA model
model	sample	bias	coverage PD	MSE	Type 1E
One-step	500	-0.03	0.98	0.07	0.35
	1000	-0.02	0.98	0.04	0.45
Two-stage	500	-0.02	0.98	0.07	0.34
	1000	-0.01	0.98	0.04	0.44
			Data generating model PE		
		PE model	Average effects		PA model
One-step	500	0.00	0.97	0.07	0.47
	1000	-0.01	0.97	0.04	0.54
Two-stage	500	0.00	0.97	0.07	0.46
	1000	-0.01	0.97	0.04	0.52

Table 4

(Average) Bias, Coverage and MSE of the Direct effect(s) of Z on Y for models PD and PE, and Type 1 error rate for $P(X|Z)$ for analyzing data generated from Models PD and PE with model PA. Uneven group-level class sizes group level Z.

age, MSE and SE/SD. The two-stage model shows similar performance to the one-stage approach. As the strength and amount of misspecification increases the bias in the PA model increases. Table 4 shows the results for models PD and PE for the uneven class size model with Z at higher level. Similarly to the even class size condition when PA model is used the Type 1 error rate is high with both estimators, but using the correct data generating model the bias is almost nonexistent with both estimators. The supplementary materials show the results for data generated with Z at the lower level. The results show the same tendencies there as well.

Application: Predicting task variety in self managing teams

We illustrate the use of the two-stage estimator based on a dataset collected by van Mierlo et al. (2005) in 5 large scale health care organizations in the Netherlands. The dataset is available as example dataset in Latent Gold (Vermunt & Magidson, 2013) and was used by Vermunt (2003) when introducing the one-stage multilevel latent class model. The LC model is based on 5 indicators measuring the perception of task variety of employees. The initial four response categories were collapsed into two. The 5 items (translated from Dutch) are (with the shorthand notation in parentheses):

- Do you always do the same things in your work [Nonrepetitive]
- Does your work require creativity? [Creativity]
- Is your work diverse? [Diverse]
- Does your work make enough usage of your skills and capacities? [Capacity]
- Is there enough variation in your work? [Variation]

Following Vermunt (2003) we present a model with two classes at both the higher and the lower level using the non parametric parametrization[1]. On the employee level the larger class, the "Diverse" (65%) class is characterized by high levels of task variation, diversity and creativity. On the other hand the "Structured" class (35%) is characterized by repetitive, not creative, and unvaried tasks. On the group level members of teams in the first group-level class (66% of teams) are most likely to belong to the first individual-level class, having more diverse tasks, involving more capacity and variation. The second group level class is comprised of teams whose members are most likely to belong to second individual level class, having more uniform, repetitive tasks.

In the next step we add covariates to the model, explaining membership in the task variety employee level classes by age, job tenure,working hours and gender (for all covariates the same re-coding is applied as in the dataset that can be found in Latent GOLD as example dataset for multilevel LCA (Vermunt & Magidson, 2013)). The estimates obtained with the one and two-stage estimators are very similar as we can see in Table 6. Tenure and working hours have a significant effect on class membership. The first cluster has the oldest workers, while the younger age groups are more likely to be associated with the second Structured class. At the same time the Structured class is associated with less working hours than the Diverse class. Gender and age are not significant predictors. After running the model with covariates we investigated the residual association

[1] Readers interested in a detailed description of model selection for this dataset can consult Nagelkerke et al. (2015), in the following we focus on the simplest model introduced by Vermunt (2003) that suffices for our example.

	Gclass 1	Gclass 2	Class 1	Class 2
	Diverse	Uniform	Diverse	Structured
Size	0.66	0.34	0.65	0.35
nonrepetitive	0.43	0.28	0.51	0.14
creative	0.61	0.44	0.71	0.27
diverse	0.80	0.49	0.97	0.20
capacity	0.74	0.58	0.83	0.42
variation	0.77	0.46	0.93	0.17
class 1	0.79	0.39	.	.
class 2	0.21	0.61	.	.

Table 5

The multilevel latent class model of task-variety

	One-step		Two-stage	
	β	SE	β	SE
age (young)	-0.20	0.09	-0.20	0.09
age (mid)	-0.07	0.07	-0.07	0.07
age (old)	0.05	0.08	0.05	0.08
tenure (low)	-0.18	0.08	-0.17	0.08
tenure (high)	-0.08	0.08	-0.08	0.08
working hours (part time)	-0.25	0.07	-0.25	0.07
working hours (full time)	0.15	0.08	0.15	0.08
gender (male)	-0.06	0.10	-0.06	0.09
		Direct effects		
age (young)	-0.20	0.09	-0.20	0.09
age (mid)	-0.07	0.07	-0.07	0.07
age (old)	0.05	0.08	0.06	0.08
tenure (low)	-0.18	0.08	-0.17	0.08
tenure (high)	-0.08	0.08	-0.08	0.08
working hours (part time)	-0.23	0.07	-0.23	0.07
working hours (full time)	0.15	0.08	0.15	0.08
gender (male)	-0.06	0.10	-0.06	0.09
working hours on capacity	-0.36	0.11	-0.36	0.11
working hours on capacity	-0.04	0.14	-0.04	0.14

Table 6

Covariate effects on the task-variety latent classes estimated using one-step and two-stage approaches for no direct effect and with direct effect on the capacity item

Dependent	nonrepetitive	creative	diverse	capacity	variation
nonrepetitive	.				
creative	1.1982	.			
diverse	0.009	4.2479	.		
capacity	0.6336	0.1291	2.1765	.	
variation	0.0461	1.0044	0.2956	1.6647	.
Independent	nonrepetitive	creative	diverse	capacity	variation
age	2.6127	1.8704	0.1405	0.3997	0.0986
tenure	3.2531	1.2789	0.001	3.0518	0.1009
working hours	1.995	1.3908	0.3431	5.388	1.0515
gender	4.1779	3.3911	0.3207	2.715	0.857
Twolevel	nonrepetitive	creative	diverse	capacity	variation
Group	1.5345	1.3146	0.6537	0.9965	0.6484
Pairs	1.3852	1.8547	0.0892	0.1948	0.0737

Table 7

Bivariate residual statistics for the model with covariate effect on the lower level classes

between the items of the LC model and the covariates using the bivariate residuals (BVR), see Table 7. As a rule of thumb values higher than 3 show evidence of some residual association[2] As working hours showed a high residual association with capacity (BVR=5.89) we allowed for a direct effect between the two. The model with direct effects is shown in the lower half of Table 6. We can see that adding the direct effect the general conclusions do not change significantly in this case. The effect of working hours on the latent classes decreases marginally, and the direct effect is significant, showing a higher effect of working hours on the first class. The overall conclusion for the rest of the model is not affected.

Discussion

We introduced a two-stage estimator of the multilevel latent class model, that separates the estimation of the measurement and structural model by fixing the measurement model parameters to values estimated at the first stage when estimating the structural model (second stage). The proposed estimator is flexible enough to allow for freeing paramaters of the measurement model while estimating the structural model where necessary.

We investigated the bias, coverage and MSE of the proposed two-stage and the alternative mainstream one-stage estimator. When all model assumptions hold the proposed two-stage estimator has similar properties to the one-stage estimator.

We investigated the bias of both estimators in conditions where a direct effect between the covariate and item(s) of the latent class model are present. The performance of the two estimators was very similar in these situations as well, namely as the severity of the underlying violations increases ignoring them leads to bias with both approaches. When analyzing the data with the correct data generating model the two-stage approach performs well.

[2] This rule of thumb is based on the assumption that the bivariate residuals follow a chi square distribution with 2 df that does not hold, yet given the complexities of approximating the distribution of the BVR statistic this rule of thumb is often used.

We generated data from models (model PD and PE) where no effect exists between the latent variable and the covariate, only direct effect(s) between the covariate and indicator(s). Analyzing these data assuming no DE, but only regressing class membership on the covariate introduces a Type 1 error rate above the nominal level. The more unmodeled direct effects are present the higher the type 1 error rate is ignoring these effects.

An issue to take into account with two-stage estimators is how to account for the uncertainty about the fixed parameters in the calculation of the stage two standard errors. Pseudo ML estimates have two sources of variability: the variability due to sampling in step two, but also that of the sampling variability of step one (Gong & Samaniego, 1981). For single level two-stage LCA models variance estimators that correct for the uncertainty due to the step 1 estimates are available (Bakk & Kuha, 2018). However simulation studies show that the correction factor is negligible for models where the measurement model is strong and the sample size large enough. As such in the current paper we ignore the variability due to the sampling variability in the step one estimates. The results show that in all conditions while the coverage is marginally lower then for the one-stage model, the difference is very small.

An alternative stage wise estimator, the bias-adjusted three-step approach has already been investigated for latent Markov models for longitudinal data that have a similar nested data structure - units nested in time points. However, while formulas to compute the classification error in such models are easy to derive based on the LM model assumptions and of the Markov properties, computation of the classification error probabilities is not as straightforward for “pure" multilevel data due to the interaction of the individual level latent variable with the group level one; in addition, the bias-adjusted three-step approach focuses only on the structural model in the third step. Thus possible misspecifications in the measurement model - like unmodeled direct effects - cannot be detected. How to extend the three-step approach to *multilevel* LC modeling can be an interesting topic for future research.

References

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

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

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling*.

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

Bakk, Z., Tekle, F. T., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, 272-311.

Cheng, S.-Y., Shen, A. C.-T., & Jonson-Reid, M. (2020). Profiles of teen dating violence and association with depression among chinese teens. *Journal of Innterpersonal Violence*.

Cole, V. T., Bauer, D. J., & Hussong, A. M. (2019). Assessing the robustness of mixture models to measurement noninvariance. *Multivariate Behavioral Research*, *54*(6), 882-905. Retrieved from `https://doi.org/10.1080/00273171.2019.1596781` (PMID: 31264477) doi: 10.1080/00273171.2019.1596781

Da Costa, L. P., & Dias, J. G. (2015). What do europeans believe to be the causes of poverty? a multilevel analysis of heterogeneity within and between countries. *Social Indicators Research*, *122*(1), 1–20.

Di Mari, R., & Bakk, Z. (2018). Mostly harmless direct effects: A comparison of different latent markov modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 467-483.

Fagginger Auer, M. F., Hickendorff, M., Van Putten, C. M., Bèguin, A. A., & Heiser, W. J. (2016). Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies. *Applied Measurement in Education*(29), 144-159.

Finch, W. H., & French, B. F. (2014). Multilevel latent class analysis: Parametric and nonparametric models. *The Journal of Experimental Education*, *82*(3), 307–333.

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 861–869.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology*, 79-259.

Hagenaars, J. A. (1990). *Categorical longitudinal data- loglinear analysis of panel, trend and cohort data*. Newbury Park, CA:Sage.

Horn, M. L. V., Fagan, A. A., Jaki, T., Brown, E. C., Hawkins, J. D., Arthur, M. W., ... Catalano, R. F. (2008). Using multilevel mixtures to evaluate intervention effects in group randomized trials. *Multivariate Behavioral Research*, *43*(2), 289-326. Retrieved from `https://doi.org/10.1080/00273170802034893` (PMID: 26765664) doi: 10.1080/00273170802034893

Hsieh, T.-C., & Yang, C. (2011). Multi-level latent class analysis of internet use pattern in taiwan. In J. J. Yonazi, E. Sedoyeka, E. Ariwa, & E. El-Qawasmeh (Eds.), *e-technologies and networks for development* (pp. 197–208). Berlin, Heidelberg: Springer Berlin Heidelberg.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*(364), 805–811.

Lanza, S. T., Rhoades, B. L., Nix, R. L., & Greenberg, M. T. (2010). Modeling the interplay of multilevel risk factors for future academic and behavior problems: A person-centered approach. *Development and Psychopathology*, *22*(2), 313-335.

Lukociene, O., Varriale, R., & Vermunt, J. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*(1), 247–283.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180-197.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA:Sage.

Morselli, D., & Glaeser, S. (2018). Economic conditions and social trust climates in europe over ten years: An ecological analysis of change. *Journal of Trust Research*, *8*(1), 68-86. Retrieved from `https://doi.org/10.1080/21515581.2018.1442722` doi: 10.1080/21515581.2018.1442722

Mutz, R., Bornmann, L., & Daniel, H.-D. (2013). Types of research output profiles: A multilevel latent class analysis of the austrian science funds final project report data. *Research Evaluation*, *22*(2), 118–133.

Mutz, R., & Daniel, H. (2013). University and student segmentation: Multilevel latent-class analysis of students' attitudes towards research methods and statistics. *British Journal of Educational Psychology*, *83*(2), 280-304.

Nagelkerke, E., Oberski, D., & Vermunt, J. (2015, 06). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, *46*. doi: 10.1177/0081175015581379

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782-797. doi: 10.1080/10705511.2016.1221313

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A monte carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*(3), 267–279.

Paccagnella, O., & Varriale, R. (2013). Asset ownership of the elderly across europe: A multilevel latent class analysis to segment countries and households. In N. Torelli, F. Pesarin, & A. Bar-Hen (Eds.), *Advances in theoretical and applied statistics* (pp. 383–393). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-642-35588-2_35` doi: 10.1007/978-3-642-35588-2_35

Rindskopf, D. (2006). Heavy alcohol use in the fighting back survey sample: Separating individual and community level influences using multilevel latent class analysis. *Journal of Drug Issues*, *36*(2), 441-462.

Ruelens, A., & Nicaise, I. (2020). Investigating a typology of trust orientations towards national and european institutions: A person-centered approach. *Social Science Research*, *87*, 102414. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0049089X20300120` doi: https://doi.org/10.1016/j.ssresearch.2020.102414

Tomczyk, S., Hanewinkel, R., & Isensee, B. (2015). Multiple substance use patterns in adolescentsa multilevel latent class analysis. *Drug and alcohol dependence*, *155*, 208–214.

van Mierlo, H., Rutte, C. G., Kompier, M. A. J., & Doorewaard, H. A. C. M. (2005). Self-managing teamwork and psychological well-being: Review of a multilevel research domain. *Group & Organization Management*, *30*(2), 211-235. Retrieved from https://doi.org/10.1177/1059601103257989 doi: 10.1177/1059601103257989

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213-239.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Vermunt, J. K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic and Advanced and Syntax*. Belmont Massachusetts: Statistical Innovations Inc.

Yu, H.-T., & Park, J. (2014). Simultaneous decision on the number of latent clusters and classes for multilevel latent class models. *Multivariate Behavioral Research*, *49*(3), 232-244. Retrieved from https://doi.org/10.1080/00273171.2014.900431 (PMID: 26735190) doi: 10.1080/00273171.2014.900431

Zhang, X., van der Lans, I., & Dagevos, H. (2012). Impacts of fast food and the food retail environment on overweight and obesity in china: a multilevel latent class cluster approach. *Public health nutrition*, *15*(1), 88–96.