Yajing Zhu, Fiona Steele and Irini Moustaki

A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome

Article (Accepted version)
(Refereed)

A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome

Yajing Zhu, Fiona Steele, Irini Moustaki

London School of Economics and Political Science

Author Note

Address for correspondence: Yajing Zhu, Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K. Email: y.zhu18@lse.ac.uk

Abstract

The 3-step approach has been recently advocated over the simultaneous 1-step approach to model a distal outcome predicted by a latent categorical variable. We generalise the 3-step approach to situations where the distal outcome is predicted by multiple and possibly associated latent categorical variables. Although the simultaneous 1-step approach has been criticised, simulation studies have found that the performance of the two approaches is similar in most situations (Bakk & Vermunt, 2016). This is consistent with our findings for a 2-LV extension when all model assumptions are satisfied. Results also indicate that under various degrees of violation of the normality and conditional independence assumption for the distal outcome and indicators, both approaches are subject to bias but the 3-step approach is less sensitive. The differences in estimates using the two approaches are illustrated in an analysis of the effects of various childhood socioeconomic circumstances on BMI at age 50.

Keywords: 3-step approach, latent class analysis, robustness, multiple latent variables

A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome

Introduction

In recent years, methodological developments have been proposed to relate a latent categorical variable to additional variables of interest in regression models where latent class membership is the outcome variable (e.g. Asparouhov & Muthén, 2014a; Vermunt, 2010) or a predictor of an outcome (e.g. Bakk, Tekle, & Vermunt, 2013; Bakk & Vermunt, 2016). Of these two extensions, the latter is of particular interest in life-course research where questions of interest include how childhood circumstances influence later life outcomes such as overall well-being. These latent, and possibly multi-dimensional, childhood circumstances may be related to a set of observed indicators or items through a measurement model and treated as predictors of an adult outcome in a regression model. In this paper we focus on situations where the outcome is temporally distal to the latent construct(s), in which case the outcome should not influence the measurement of the latent variable.

Previous research has focused on the case of a single latent variable C, measured by p observed responses $\mathbf{U} = (U_1, \dots U_p)$, and a distal outcome Z. In general, methods for this situation can be categorised into two major groups: the 1-step approach and various step-wise approaches(including the modal class, modified Bolck-Croon-Hagennars (BCH), Lanza-Tan-Bray (LTB) and 3-step maximum likelihood methods). The 1-step approach simultaneously estimates the measurement model and the regression model of Z on C, treating Z as an additional indicator for C. Parameter estimates, including item response probabilities and regression coefficients, are obtained by jointly maximizing the log-likelihood of response patterns and the distal outcome. There are three main advantages of the 1-step approach. First, it is more efficient compared to step-wise approaches that may introduce additional uncertainty between steps; second it allows for more flexible model structures, such as models with direct effects of covariates on indicators and the distal outcome; and, third, it is straightforward to account for residual correlation between Z and Us, beyond that captured by class membership (Bakk et al., 2013). However, the 1-step approach has received criticism over the past ten years regarding the practicality of the simultaneous estimation and the requirement for additional distributional assumptions about Z. Vermunt (2010) noted the burden of having to re-estimate the entire model should one decide to add or delete covariates in the measurement model. He also pointed out a more serious issue that the inclusion of a distal outcome into the measurement model creates an unintended circular relationship in that the latent class C that is supposed to explain Z is also determined partly by Z (also discussed in Bakk & Vermunt, 2016). If there are multiple distal outcomes, the shift in the latent class proportions can be severe, especially when the classes outnumber the indicators or when class separation is poor. Moreover, by treating Z as an indicator for C, the 1-step approach requires additional assumptions and, for continuous Z, that Z is normally distributed within classes (Asparouhov & Muthén, 2014a; Bakk et al., 2013; Bakk & Vermunt, 2016).

When there is more than one latent categorical variable, extensions to standard step-wise approaches are required. The consideration of such an extension was first suggested by Bolck et al. (2004). Bakk et al. (2013) also discussed briefly an application of the 3-step maximum likelihood (ML) approach with two latent categorical variables where one predicts the other. Our proposed model has a more flexible specification that allows for an association between the latent variables through a log-linear model. Similar structures can be found in the structural equation modelling literature. For example, B. Muthén (2001) discussed a confirmatory latent class analysis of a two-wave panel study, with two associated latent class variables for antisocial behaviour; and mixture growth modelling with repeated measures of two related latent variables capturing fundamental individual differences, where each class has a unique set of growth parameters. In social research, it is common to have more than one latent predictor and researchers may wish to treat these as categorical (e.g. socio-economic situations). In the structural equation modelling framework, Ploubidis et al. (2015) considered an application with two latent variables, with a causal relationship, that jointly predict distal outcomes (using the modal class approach). Bauldry et al. (2016) fitted a model that estimates the effects of two associated continuous latent summaries of perceptions of physical and personality attractiveness on education attainment (using the 1-step approach).

This article contributes to the existing literature in several ways. First, we propose an extended 3-step approach that can relate more than one, possibly associated, latent categorical variables to a distal outcome. Second, building upon the recent investigations of Asparouhov and Muthén (2014a) and Bakk and Vermunt (2016), we further examine the robustness of the general 3-step approach by considering additional forms of non-normality in the distribution of Z within classes, such as distributions with skewness and excess kurtosis. In addition, our work fills a gap in current research by comparing the 3-step approach with the 1-step approach. Although the limitations of the 1-step approach are now well-established, few studies have evaluated the relative robustness of the 1-step and 3-step approaches (Asparouhov & Muthén, 2014a, 2014b). Third, we assess the impact of the violation of the common assumption of conditional independence between Z and the Us for both the 1-step and 3-step ML approaches. In addition, extending the work of Asparouhov and Muthén (2014a, 2014b) which shows that the 1-step approach leads to a shift in the latent class proportions when Z follows a bimodal distribution within classes, we further explore whether the number of classes needed to capture the association among the Us is altered. This has not been discussed before and is more serious because if the number of classes changes, the interpretation of the latent variable may no longer hold.

The rest of the paper is organised as follows. We start with a brief review of the step-wise methods used for situations with a single latent categorical variable and a distal outcome. Next, we propose a general 3-step ML approach for multiple latent variables and demonstrate its performance through both simulations and an empirical example. We conclude with a further discussion and recommendations for future research.

Review of step-wise methods

Various step-wise approaches have been proposed in recent years in order to preserve the latent classes from the measurement model for U. The key difference from the 1-step approach is that in step-wise approaches the measurement model is estimated separately, with parameters from this step carried forward in later analyses that involve external variables. Commonly used approaches are the modal class approach, the modified Bolck, Croon and Hagenaars (BCH) approach (Vermunt, 2010, developed from Bolck et al., 2004), the 3-step ML approach (with modal or proportional assignments) and the Lanza-Tan-Bray (LTB) approach (Lanza, Tan, & Bray, 2013). In the following, we summarise the key concepts and restrictions of each method in situations where the latent variable predicts the distal outcome. The methods are described for a single latent variable, as in previous research. We consider the extension to multiple latent variables in the next section.

The modal class approach

After estimating the latent class model, the modal class assignment (M) is saved for each individual based on their posterior probability of being in each class, i.e. $P(C|\mathbf{U})$. The modal class membership is often treated as a known nominal covariate in further analysis. However, this approach ignores the uncertainty of classification, i.e. the probability of a class assignment r given the true class k (misclassification error):

$$\begin{aligned} P(M=r|C=k) &= \sum_u P(M=r,\mathbf{U}=u|C=k) \\ &= \frac{\sum_u P(C=k|\mathbf{U}=u)P(M=r|\mathbf{U}=u)P(\mathbf{U}=u)}{P(C=k)}, \end{aligned} \tag{1}$$

where r, $k \in \{1,\ldots,K\}$ index the respective modal and latent classes. $P(M|C)$ is therefore a result of averaging the misclassification error over all patterns of U. Using M in place of

C can therefore lead to biased estimates and spurious statistical inferences of class effects (and of the effects of covariates correlated with the latent variable) on the distal outcome. A modification of the modal class approach is called the pseudo class approach (Clark & Muthén, 2009). Instead of assigning individuals to classes with certainty, individuals are now assigned to classes randomly sampled from the multinomial distribution based on the posterior probability of being in each class. However, when the class separation is poor, we can expect a large amount of classification error in both methods that may lead to biased estimates for coefficients of Z on C.

When it comes to modelling the relationship between Z and M, one needs to take into account that (Z, M) is different from (Z, C) and therefore a correction is needed.

The modified BCH approach

In this paper we are interested in studying the effects of latent predictors on a distal outcome, i.e. the conditional distribution of Z given C. To achieve this, one needs to first establish the relationship between (Z, M) and (Z, C). Translating equation (9) of Bolck et al. (2004) into our context, we have

$$P(M = r, Z = z) = \sum_k P(C = k, Z = z)P(M = r|C = k), \tag{2}$$

which can be expressed in terms of the conditional relationship of Z given C as

$$P(M = r, Z = z) = \sum_k P(Z = z|C = k)P(C = k)P(M = r|C = k), \tag{3}$$

which is equivalent to a latent class model with two indicators (M, Z). Equations (2) and (3) also imply that to obtain $P(Z = z|C = k)$, one needs to adjust for the misclassification probability $P(M = r|C = k)$ given by (1). Also worth noting is that to obtain (2) and (3), it is implied that M depends only on U (as M is only determined in the measurement model) and we assume the Us are conditionally independent of Z given C.

The modified BCH approach (Vermunt, 2010) originated from the BCH approach proposed by Bolck et al. (2004). It was first developed for latent class models with covariates before being extended by Bakk et al. (2013) to the situation where Z is a distal outcome. The weighted pseudo log-likelihood function is:

$$l = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \text{log} P(Z = z_i | C = k) P(C = k), \tag{4}$$

where i indexes subjects, $w_{ik} = \sum_{r=1}^{K} p_{ir} d_{rk}$ and $p_{ir} = P(M = r | \mathbf{U} = u_i)$ for $k, r \in \{1, \ldots, K\}$. Note that d_{rk} represents an element of the inverted $K \times K$ matrix D of the misclassification probabilities $P(M = r | C = k)$. A detailed description of this approach is available in Vermunt (2010) and Bakk et al. (2013).

Using the modified BCH approach, the latent class solution derived from Step 1 remains unchanged. It is also robust to violation of the assumptions that Z is normally distributed within classes and has constant variance across classes (Asparouhov & Muthén, 2014b; Bakk & Vermunt, 2016). However, estimation problems may arise for categorical Z when negative cell frequencies are obtained in the weighted cross-classification of categorical Z and C (Asparouhov & Muthén, 2014b; Bakk et al., 2013). This problem may be exacerbated when there are multiple distal outcomes. Bakk, Oberski, and Vermunt (2016) also show that standard errors are underestimated when the sample size is small and class separation is poor.

The Lanza-Tan-Bray approach

The LTB approach was first proposed by Lanza et al. (2013) as a method that preserves the latent class solution estimated from the measurement model. To estimate the class-specific means for outcome Z, they first estimate $P(C|Z)$, treating Z as a predictor of C. Bayes$'$ theorem is then applied to obtain $P(Z|C)$ by using the kernel density to approximate $P(Z)$. In a later modification by Asparouhov and Muthén (2014a), the sample distribution of Z is used which produces similar results to using the kernel approximation. However, it suffers from several limitations. First, the measurement model cannot include covariates as otherwise both the covariates and the outcomes will be predicted by C after Bayes$'$ transformation (Asparouhov & Muthén, 2014a). Second, when the within-class distribution of Z has outliers, the estimated mean of Z is severely biased Bakk, Oberski, and Vermunt (2014). They also find that for continuous Z the LTB approach produces heavy bias when heterogeneity across classes is not accounted for. A further limitation of the LTB approach is that it cannot be generalised to multiple distal outcomes, unless they are strictly conditionally independent given C. Further details about the LTB approach, including standard error estimation can be found in Asparouhov and Muthén (2014a) and Bakk et al. (2016).

The 3-step ML approach

The 3-step ML approach is based on the idea of the modified BCH approach, which was first proposed by Vermunt (2010). To illustrate this method, consider a simple model with one latent categorical variable C measured by observed indicators $\mathbf{U} = (U_1, ..., U_p)$ and a binary distal outcome Z predicted by both C and an observed covariate X. Parameter estimation involves the following three steps (also illustrated in Figure 1a). Note that Step 1 and 2 of the modified BCH and the 3-step ML approaches are essentially the same; only Step 3 differs. Further details are available in Vermunt (2010), Bakk et al. (2013) and Asparouhov and Muthén (2014a).

Step 1: Perform a latent class analysis (LCA) without Z or X. Calculate the posterior probability of being in each class and the modal class M for each individual.

Step 2: Calculate the misclassification probabilities given in (1), which will be treated as fixed quantities in Step 3.

Step 3: The log-likelihood function is given as:

$$\log L_{3step} = \sum_{i=1}^{N} \log \sum_{k=1}^{K} P(Z = z_i | C = k, X) P(C = k) P(M = r | C = k). \tag{5}$$

The key advantage of the 3-step ML approach is that it preserves the class solution in the measurement model and that the efficiency of the 3-step ML approach is close to 1-step approach. Similar to the modified BCH method, the misclassification probabilities obtained from Step 1 are carried forward in subsequent analyses. One limitation is that although Step 1 estimates the measurement model separately from Z, the inclusion of Z in Step 3 may lead to a change in the latent class proportions when the within-class distribution of Z is bimodal (Asparouhov & Muthén, 2014a; Bakk & Vermunt, 2016). The 3-step ML approach also understates the standard errors by treating the misclassification probabilities in Step 2 as observed, rather than estimated from Step 1. A standard error correction method was proposed by Bakk et al. (2014) that takes into account this additional source of variation.

Comparison of methods: Evidence from simulation studies

A number of simulation studies have been conducted to compare the performance of different methods in a range of situations including varying entropy levels and sample sizes. Vermunt (2010) considered a latent class model with covariates that predict class membership. He finds that the modified BCH and 3-step ML approaches result in slightly downward-biased estimates, while 1-step estimates have a slight upward bias (averaging across all scenarios with varying entropy levels and sample sizes). It has also been noted that when the sample size is small and entropy is low (0.36), the 1-step, the modified BCH, and the 3-step ML approaches all fail, although estimates from the 1-step approach are less biased than those from the latter two approaches, especially for large samples (N=10,000). One possible explanation is that at low entropy levels, the differences between classes are over-stated, which leads to an underestimation of the classification error (Bakk et al., 2014; Vermunt, 2010). Standard errors are severely underestimated using the modified BCH approach, although using a sandwich variance estimator provides a slight correction. Both the 3-step ML and 1-step approaches give average estimated standard errors that are close to the standard deviation of the parameter estimates across replications; the former SE is slightly underestimated while the latter is slightly overestimated. The 3-step ML approach is also shown to be roughly as efficient as the 1-step approach.

From studies that considered latent class models with latent variables as predictors of a distal outcome, we can conclude the following. When all necessary model assumptions hold, the sample size is large and class separation is good (entropy>0.6), all methods perform well with small bias, correct SEs and good coverage. When the sample size is small and entropy is low (<0.6), all methods can fail with either large bias or poor coverage (Asparouhov & Muthén, 2014a, 2014b; Bakk et al., 2013; Lanza et al., 2013), although the 1-step approach slightly outperforms other methods (Asparouhov & Muthén, 2014a, 2014b).

The robustness of each method to departures from normality has been investigated for continuous Z. When Z follows a bimodal distribution within classes, the class proportions may be affected for both the 1-step and 3-step ML approaches, which then leads to heavily biased estimates of the effects of C on Z (Asparouhov & Muthén, 2014a, 2014b; Bakk & Vermunt, 2016). It has also been noted by Asparouhov and Muthén (2014a) that when estimates for class proportions from Step 1 are used as starting values in Step 3 (instead of random sets of starting values), the latent class solution remains unchanged. However, in a later investigation of the impact of a bimodal Z, the class proportions from Step 1 and Step 3 differ significantly in almost all replications when entropy is 0.7 (Asparouhov & Muthén, 2014b). The modified BCH approach provides unbiased estimates but poor coverage (around 89%) across all entropy levels, particularly when entropy is low. This is mainly because the weights w_{ik} depend on the misclassification error, which has a higher variability when class separation is unclear (Asparouhov & Muthén, 2014b). When Z has a medium or low degree of bimodality, the LTB approach results in larger bias when class separation is low and unbiased estimates coupled with poorer coverage when class separation is high, compared with the 3-step ML approach that allows for unequal class-specific residual variances (Asparouhov & Muthén, 2014b; Bakk & Vermunt, 2016). Bakk and Vermunt (2016) also showed that both the modified BCH and the 3-step ML approaches are insensitive to unequal class-specific variances because they explicitly allow for unequal variances, while the LTB approach produces large bias.

Considering the robustness, efficiency, interpretability and the potential for generalisation to more complex model structures with possibly mixed types of distal outcomes, the 3-step ML approach is particularly appealing. Further investigations of this approach are therefore the focus of this research.

A GENERAL 3-STEP ML APPROACH FOR MULTIPLE LATENT VARIABLES

We now consider an extension of the 3-step ML method for models with multiple latent categorical variables. The method is described for two latent categorical variables but extensions to include multiple categorical variables are straightforward, and a model with four latent variables is considered in the empirical study. Figure 1b illustrates the structure of this model, where C_1 and C_2 denote two latent categorical variables (LV) and M_1 and M_2 are the corresponding modal classes derived from separate latent class analyses of two distinct sets of indicators (not shown in Figure 1b for simplicity). The associated posterior probabilities of being in each class, as well as the misclassification probabilities, are calculated in the second step for use in the last step. The curved arrow between the two latent variables indicates the existence of an association. For a general setting (without specifying the direction of the association between C_1 and C_2), a log-linear model can be specified. As the latent variables are both categorical, we consider the cross-classification of C_1 and C_2 assuming two classes for each latent variable. Let k_1 and k_2 be the class index for each latent variable (k_1, $k_2 \in 1$, 2) and $\mu_{k_1 k_2}$ the expected frequency in each cell. We assume cell counts $\sim$ i.i.d Poisson($\mu_{k_1 k_2}$).

A log-linear model with a two-way interaction between C_1 and C_2 can be specified as

$$\log(\mu_{k_1 k_2}) = \omega_0 + \omega_{k_1}^{C_1} + \omega_{k_2}^{C_2} + \omega_{k_1 k_2}^{C_1 C_2}, \tag{6}$$

where ω_0 is the intercept term, $\omega_{k_1}^{C_1}$ and $\omega_{k_2}^{C_2}$ are the main effects of latent variables C_1 and C_2 and $\omega_{k_1 k_2}^{C_1 C_2}$ is their interaction effect. This is the saturated model for the 2×2 table of cell frequencies. As the marginal frequencies in the cross-classification are known, we only have four free cell frequencies to estimate, but (6) includes nine. Thus for model identification, the following five constraints are imposed,

$$\omega_2^{C_1} = \omega_2^{C_2} = \omega_{12}^{C_1C_2} = \omega_{21}^{C_1C_2} = \omega_{22}^{C_1C_2} = 0, \tag{7}$$

where category 2 is taken as the reference for each latent variable. The ωs can be interpreted as log-odds or log-odds ratios:

$$\omega_1^{C_1} = \log\Big(\frac{\mu_{12}}{\mu_{22}}\Big),\ \omega_1^{C_2} = \log\Big(\frac{\mu_{21}}{\mu_{22}}\Big),\ \omega_{11}^{C_1C_2} = \log\Big(\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}}\Big), \tag{8}$$

where $\omega_{11}^{C_1C_2}$ indicates the association between C_1 and C_2.

After specifying the association between the latent variables, the last step is to fit a regression model to estimate the effects of C_1, C_2 and X on distal outcome Z. For example, a logit model for binary Z can be written:

$$\text{logit}[P(Z = 1|X, C_1 = k_1, C_2 = k_2)] = \tau_0 + \tau_{k_1}^{C_1} + \tau_{k_2}^{C_2} + \lambda X, \tag{9}$$

where $\tau_{k_1}^{C_1}$ and $\tau_{k_2}^{C_2}$ denote the main effects of C_1 and C_2. For the ease of illustration, we assume the effect of X, λ, is constant across classes of C_1 and C_2 and that interaction effects are not considered, although it is straightforward to extend the model to include interactions between the latent variables or between a latent variable and X.

Similar to (5), the log-likelihood of the observed data for a random sample of size N is:

$$\log L_{\text{2-lv3step}} = \sum_{i=1}^{N} \log \sum_{C_2} \sum_{C_1} P(C_1, C_2) P(Z = z_i|X, C_1, C_2) P(M_1, M_2|C_1, C_2). \tag{10}$$

Note that in Step 1, the latent class models are estimated separately (to ensure the class solution of C_1 is not determined by indicators for C_2) and the association between two latent variables is only introduced in Step 3. We can therefore simplify the third component of (10) to

$$P(M_1, M_2|C_1, C_2) = P(M_1|C_1)P(M_2|C_2), \tag{11}$$

where both terms on the right-hand side are essentially misclassification probabilities given in (1) and calculated in Step 1. $P(C_1, C_2)$ in (10) can be derived from the log-linear model for latent variables using (6), where $P(C_1 = k_1, C_2 = k_2) = \mu_{k_1 k_2}/N$.

The extension of the 3-step ML approach to include multiple categorical latent variables allows for a flexible specification of the association between latent variables. However, we should also note that the above misclassification probabilities are fixed and treated as known. As they are in fact calculated from the LCA in Step 1, the standard errors of estimates in the 3-step ML method are expected to be understated. Bakk et al. (2014) also highlighted potential sources of bias, for instance, if the within-class distribution of continuous Z is bimodal or the error variance is not constant across classes.

SIMULATION STUDY

As noted above, the few studies that have compared the performance of the 3-step ML and simultaneous 1-step methods have found that their performance is often similar. We build upon earlier work by conducting a simulation study to further investigate the relative performance of the two methods under departures from two key model assumptions (within-class normality of Z and conditional independence of Us and Z) and for the extension to two categorical latent variables. For both investigations, we are concerned with potential bias of coefficients for the latent variables in the model for Z, as well as a more fundamental problem where the number of classes that are needed to capture the association among the Us may be altered. Extending the work of Bakk and Vermunt (2016) that examined the robustness of the 3-step ML approach for bimodal and heterogeneous class-specific distributions of Z, we consider the performance of the general 3-step ML approach and the 1-step approach under other forms of non-normality, i.e. skewness and excess kurtosis. These two forms of non-normality are common in practice and may not be well captured by a finite mixture of normal components. Previous research has found that non-normality of Z can lead to both biased coefficients and shifted class proportions for both methods (Asparouhov & Muthén, 2014a; Bakk & Vermunt, 2016). We build on the literature by investigating whether non-normality of Z affects the number of classes needed in the measurement model for the 1-step approach. We anticipate that additional (spurious) classes may be required to capture the distribution of Z. The second investigation of the impact of local dependence of Z on both methods has not be considered in previous research. We anticipate that if such dependence is not accounted for, both methods can give biased estimates and the 1-step approach may even identify spurious classes.

Data generation

We generate data from models with a distal outcome Z that is predicted by two associated latent categorical variables, each measured by five binary items and with two classes. Both continuous and binary Z are considered. For simplicity, we do not include covariates but extensions are straightforward. We compare the performance of both methods in a number of scenarios where all assumptions are met (Study 1), when the normality assumption about a continuous Z is violated in various ways (Study 2) and when the conditional independence assumption is violated (Study 3). Note that assumptions considered in studies 2 and 3 are common to both methods. We generate data from a measurement model with ten dichotomous indicators, where the first five measure C_1 and the latter five measure C_2. Taking the second class as the reference for both latent variables, class 1 (2) of C_1 and C_2 gives high (low) response probabilities for all five indicators. Latent variables are generated from the log-linear model specified in (6).

Similar to previous research (Asparouhov & Muthén, 2014a; Bakk et al., 2014; Bakk & Vermunt, 2016), varying sample sizes are considered and we manipulate the entropy levels through class-specific thresholds in the measurement model. Specifically, entropy values of 0.7 (high) and 0.4 (low) correspond to logit thresholds of 1.25 and 0.75, respectively.

The distal outcome Z is generated from the model $Z = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \epsilon$ where $\epsilon \sim N(0,1)$ and from the model $\text{logit}[P(Z = 1|C_1, C_2)] = \beta_0 + \beta_1 C_1 + \beta_2 C_2$ for a binary Z. Other parameter values for the population model specific to each scenario are provided in the corresponding section. In the 3-step ML approach, as results from previous studies show that modal and proportional assignments of individuals to classes lead to similar parameter estimates in Step 3 (Bakk et al., 2016, 2013; Vermunt, 2010), and that the proportion of misclassified observations is smaller using the modal assignment, we use modal assignments in Step 1. In order to mimic empirical studies, in the 1-step approach we use 100 sets of random starting values and impose parameter constraints (greater or less than zero) on log-linear parameters for latent class allocations. These constraints are mainly set to avoid potential label switching in the class allocation of mixture models. We generate 500 replications in each study. The latent class models for the two sets of binary items are estimated separately in Mplus 7.31 (L. Muthén & Muthén, 2015); the modal class assignments and misclassification probabilities are then exported to Latent GOLD 5.0 (Vermunt & Magidson, 2013) for Step 3 of the estimation procedure. The reported summary statistics are relative bias (%), average standard error across replications (SE), standard deviation of estimates across replications (SD) and 95% coverage rates. The codes for selected simulation studies are included in the supplementary material.

Study 1: all model assumptions are satisfied

Simulations are carried out for combinations of low and high entropy levels, sample sizes of 500, 2000 and 10,000, and for correlated and independent latent variables. Due to space constraints, only the results for the more common case of correlated C_1 and C_2 are presented. As results are fairly similar for all values of N, we present results only for N=2000. For continuous Z, coefficients are set at $\beta_0 = 3$, $\beta_1 = 2$, $\beta_2 = -1.5$; for binary Z, $\beta_0 = 1.2$, $\beta_1 = 1$, $\beta_2 = -1.5$. Parameters of the log-linear model for the latent variables are set at $\omega_1 = 0.7$, $\omega_2 = 0.7$, $\omega_{12} = 0.5$, indicating a positive association between C_1 and C_2.

Similar to the results obtained for the 3-step ML approach with one latent variable (Asparouhov & Muthén, 2014a; L. Muthén & Muthén, 2015), Table 1 shows that both the 3-step ML and 1-step approaches give unbiased estimates and excellent coverage for almost all parameters when model assumptions hold. The 3-step ML approach gives a slightly lower coverage for ω_2 due to its underestimated SE. In the following studies, we investigate the relative performance of these two methods in scenarios where the model assumptions are violated in various ways.

Study 2: violation of the normality assumption about Z

In this study, we are mainly concerned with the situation where the distal outcome Z is non-normal, but we fit a standard finite mixture model assuming within-class normality. The normality assumption is common to both the 1-step and 3-step ML approaches. In the simultaneous 1-step approach, Z is treated as an additional indicator for the latent variables. We therefore hypothesise that compared to the 3-step ML approach where the measurement model in Step 1 is estimated separately from Z, the 1-step approach is more sensitive to non-normal within-class distributions. Through the simulation study, we evaluate the relative performance of the two methods when the within-class distribution of a continuous Z exhibits skewness, excess kurtosis, and bimodality, respectively. The results for bimodality are given in the supplementary material as they are similar to results for the single latent variable case (Bakk et al., 2014; Bakk & Vermunt, 2016) but with slightly lower coverage in situations with poor class separation. We also conduct simulations using the modified BCH approach (detailed results in the supplementary material) as previous research confirmed its robustness to violations of distributional assumptions (Asparouhov & Muthén, 2014b; Bakk & Vermunt, 2016).

We simulate non-normality by generating Z from a mixture of non-normal and normal distributions. We first focus on the impact of these forms of non-normality on the main coefficients of interest, i.e. β_1, β_2 (estimates for ωs are shown in the supplementary materials). Next we investigate whether the number of classes needed in the mixture model can be influenced by non-normality. For each form of non-normality, entropy is fixed at the same value for each latent variable: 0.7 (high) and 0.4 (low). Parameters of the log-linear model for latent variables are set at $\omega_1 = 0.7$, $\omega_2 = 0.7$, $\omega_{12} = -0.5$, leading to the following proportions for cells in the cross-classification of C_1 and C_2 (hereafter referred to as class patterns): 0.33 for [C_1=1, C_2=1], 0.27 for [1,2] and [2,1] and 0.13 for [2,2]. Across all sub-studies, we generate sample sizes of N=200 and 2000, where 200 can be regarded as a (very) small sample.

Study 2a: Excess kurtosis. In this scenario, we generate Z from the model $Z = 3 + 2C_1 - 1.5C_2 + \epsilon$, where ϵ is drawn from a student-t distribution with 7 degrees of freedom (excess kurtosis =2) for class patterns [1 2] and [2 1], but from $N(0, 1)$ for the other two class patterns. Selected results are summarised in Table 2. We find that the 3-step ML approach provides unbiased estimates in all situations apart from small N and low entropy, where there is an obvious change in the class proportions from Step 1 to Step 3 (see supplementary material Table 2 for estimates of ωs). For both the 3-step ML and 1-step approaches, performance is best for large N and clear class separation. Comparing the results with the BCH approach in the supplementary material, the performance of the two approaches is similar across all scenarios investigated.

Study 2b: Skewness. We now consider Z with a skewed distribution for class pattern [1 2] and [2 1] but a normal distribution for the other class patterns. In the simulation, we generate residual ϵ from the log-normal distribution with zero mean. We generate a right-skewed Z from the model $Z = 3 + 2C_1 - 1.5C_2 + \epsilon$ (skewness=5.0) for class pattern [1 2] and for class pattern [2 1], we generate a left-skewed Z from the model $Z = 3 + 2C_1 - 1.5C_2 - \epsilon$ (skewness=-5.0). For other class patterns, $\epsilon \sim N(0, 1)$. To capture the heterogeneity of the data, it is standard practice to fit a finite mixture model assuming within-class normality. The results are presented in Table 3. In general, the 3-step ML approach clearly outperforms the 1-step approach when the within-class distribution of Z

is skewed, although both approaches give biased estimates. When class separation is clear the general 3-step ML approach produces estimates with a relative bias slightly over 5%, even when the sample size is small. However, when class separation is poor, class assignment in Step 1 is shifted in Step 3 (see supplementary material Table 3 for the heavily biased estimates of the ω parameters), which can partly explain the biased coefficient estimates. It should also be noted that for the general 3-step ML approach alone, skewness seems to lead to heavier bias than bimodality and excess kurtosis. Clark and Muthén (2009) showed that kurtosis can be approximated as a quadratic function of skewness and hence if the distribution of the data is highly skewed, it also has severe excess kurtosis, which can exacerbate the bias in parameter estimates. Comparing results with the BCH approach (see supplementary material Tables 4 to 7), the BCH approach outperforms the 3-step ML approach in terms of relative bias and coverage of the estimated coefficients for all combinations of entropy levels and sample sizes tested, with the exception of the case where the class separation is poor and N=200.

Number of classes. In the above simulations, we observe that in some situations the 1-step approach produces heavily biased estimates (including those for the ω parameters shown in the supplementary material). As noted earlier, it is also possible that for the 1-step approach non-normality of Z may lead to an increase in the number of latent classes needed to fit the data (Bauer, 2007). This is a major concern in empirical studies because a change in the number of classes may alter their interpretation and hence bias their estimated effects on the distal outcome. For the 3-step ML approach, the measurement model is estimated without Z in Step 1 such that the true number of classes should be obtained (provided the assumptions in Step 1 are met).

To investigate the impact of non-normality on the number of classes required in the 1-step approach, data are generated from a model with only one latent categorical variable (1-LV) with four classes. The 2-LV model with two independent classes for each variable can be viewed as a 1-LV model with four classes. Linking with previous scenarios, the same class proportions are generated, corresponding to four class patterns in the 2-LV model but with no correlation. For data generated from a 4-class model, models with 3-5 classes are fitted and the sample size adjusted BIC (ssaBIC) and p-values from the bootstrap likelihood ratio test (BLRT) are obtained for each, following the recommendation of Nylund, Asparouhov, and Muthén (2007). All parameter values in the population model for each scenario remain the same as in the 2-LV model.

Table 4 shows the percentage of replications for which each model has the minimum BIC value and for which each model is rejected or not based on the BLRT. If the number of classes is unaltered, we expect to obtain ssaBIC close to 100% for the 4-class model and close to 0% for the 3 and 5-class models. We expect BLRT p-values (averaged across replications) close to 95% for tests of 3-class vs 2-class and 4-class vs 3-class models and close to 5% for the test of the 5-class vs 4-class model.

Of the three types of within-class non-normality of Z considered, skewness is particularly troublesome as the selection rate based on ssaBIC is the lowest (below 80%). When N=2000, we also observe that ssaBIC identifies more classes than there truly exist (ssaBIC agrees with BLRT when class separation is poor). When the sample size is small (N=200), we focus on BLRT as Nylund et al. (2007) show that ssaBIC performs poorly in such situations. When class separation is poor (low entropy), fewer classes are needed for all types of non-normality (according to BLRT). However, when the class separation is good, regardless of the sample size, none of the three types of non-normality of Z influence the ability of the 1-step approach to correctly identify the number of classes.

Combining the results from investigations of bias in the coefficient estimates and of the ability to extract the true number of classes when the within-class normality assumption is violated it is obvious that, although the assumption is common to both the 1-step and the general 3-step ML methods, the 1-step approach is more sensitive than the 3-step ML approach to all forms of non-normality. Both methods perform the worst when there is within-class skewness in Z, but the 1-step approach is also likely to alter the number of classes needed to fit the data (consistent with the findings of Bauer, 2007). Moreover, both methods perform poorly when sample size is small and the class separation is poor, as expected.

Study 3: violation of the conditional independence assumption about Z

This study investigates the impact on the number of classes extracted and the parameter estimates of violation of the assumption that Z and the Us are conditionally independent given the latent variable C. This has not been studied in previous research but the conditional independence assumption is common to both the 1-step and 3-step ML approaches, i.e. $P(U, Z|C) = P(U|C)P(Z|C)$ (see also Bakk et al., 2013). Compared to the 1-step approach, one obvious advantage of the 3-step ML approach is that it is not subject to the change in the number of classes when local dependence of Z is present as the decision of the number of classes to be retained is made in Step 1, without Z. We also anticipate that if such residual dependence is not accounted for, the 1-step approach will produce more biased estimates for the relationship between latent categorical variables and Z than the 3-step ML approach, as a wrong model with insufficient/extra classes will be estimated. Study 3 investigates the relative performance of both approaches for different entropy levels and sample sizes.

For ease of illustration, we consider a 1-LV model with four classes and continuous Z. Data were generated from a model with class proportions of 0.30, 0.25, 0.25, 0.20. We then generate ten binary indicators conditional on class membership from a logit measurement model. We consider the same high and low entropy situations by manipulating the thresholds in the measurement model described earlier, for sample sizes N=500 and 2000. We increase the small sample size from 200 (used in Study 2) because we have reformulated the 2-LV model with two classes for each variable as a 1-LV model with four independent classes. The larger sample size of N=500 helps to avoid boundary solutions due to small classes. Next, to induce local dependence between item U_{10} and Z, we introduce an additional continuous random variable $u \sim N(0, 4)$. Note that conditional independence between all Us is still valid so that the measurement model is correctly specified. In addition, we set the class-specific variance of ϵ to be 4, 3, 2, 1 for C=1, 2, 3, 4, respectively and the corresponding class-specific means are 3.5 (β_1), 5 (β_2), 1.5 (β_3) and 3 (β_4). The data are then analysed using both the 1-step and the 3-step ML approaches using the DU3STEP command in Mplus. Note that we employ a slightly different parameterisation to that in earlier simulations as we are restricted by the technicalities of the program. The parameters estimated are means of Z in each latent class rather than contrasts with a reference category.

We first check if the number of classes needed is altered in this scenario using the 1-step approach Table 6. The results show that when conditional independence holds the 1-step approach tends to identify fewer classes when the class separation is unclear and especially when sample size is small; these findings are consistent with those of Nylund et al. (2007). Second, both ssaBIC and BLRT indicate that even when the conditional independence assumption between Z and items in the measurement model is violated for only one item, there is a tendency to extract additional (spurious) classes, irrespective of the level of class separation. When there is local dependence between Z and an indicator, the percentage of times that ssaBIC favours the (K + 1)-class model over the correct K-class model increases. For example, in the high entropy case with N=500, ssaBIC suggests a correct 4-class model in 89% of replications when conditional independence holds, which decreases to 61% when the assumption breaks down. Similarly, the percentage of times that ssaBIC suggests the 5-class model increases from 13% when conditional independence holds, to 39% when the assumption breaks down.

In addition, we observe that when class separation is unclear or sample size is small, there is greater disagreement between the ssaBIC and BLRT statistics. Our findings are particularly worrying for empirical studies as the assumption of conditional independence between the distal outcome and items that measure the latent variable is rather strong. If such local dependence is not accounted for in the model, we anticipate that the problem discussed above will be exacerbated when Z is correlated with more than one item.

We next examine the impact of departures from conditional dependence on the estimated coefficients when a model with the correct number of classes is fitted using both the 1-step and 3-step ML approaches. The simulations are performed in Mplus using the DU3STEP command that allows for unequal class-specific variances. The results are reported in Table 5. For illustrative purposes, we only present results for parameters β_3 and β_4 (class-specific means of Z for $C = 3$ and $C = 4$) as they are the most biased among all βs.

Clearly, regardless of the entropy level, the 1-step approach is more sensitive to the violation of the conditional independence assumption than the 3-step ML approach; the latter produces around 10% relative bias at most. The 1-step method performs particularly poorly when entropy is low. This is expected as forcing the 1-step approach to estimate a 4-class model (when the 5-class model is a better fit) leads to changes in the interpretation of classes, resulting in larger bias in the estimates. The same rationale also applies to the observation that for the 3-step ML approach, the relative bias does not seem to reduce when the sample size increases, and regardless of the entropy level. Comparing with the results from the BCH approach (see supplementary material), we find a similar performance for 3-step ML and BCH, except for the scenario with poor class separation and large sample size, where the modified BCH approach produces greater bias.

EMPIRICAL EXAMPLE

We now illustrate the general 3-step ML approach in an analysis of the effects of four latent categorical variables, capturing different aspects of childhood socio-economic situations (SES), on body mass index (BMI) at age 50. The data are taken from the 1958 British National Child Development Study (NCDS) (Power & Elliott, 2006), a cohort study that contains four waves of childhood information (at ages 0, 7, 11 and 16). The distal outcome Z is log-transformed to adjust for positive skewness. We consider repeated measures of four aspects of childhood SES: social class (father or male head's occupation), financial difficulty, material hardship and family structure. The choice of the definition of indicators follows the work of Hobcraft (1998), Schoon, Sacker, and Bartley (2003) and Chandola, Clarke, Morris, and Blane (2006). In Step 1, we estimate a separate latent class model for each of the four sets of repeated measures. Based on the Bayesian Information Criteria (BIC), log-likelihood and bivariate residuals among indicators, we conclude that 3-class, 2-class, 3-class and 2-class models best fit the data for the four dimensions of SES. Table 8 of the supplementary material summarises the derived modal class membership from the latent class analyses. Note that the labels of classes are assigned by examining the pattern of estimated response probabilities conditional on latent class membership over time. The Latent GOLD code for this empirical study is included in the Appendix.

For ease of illustration the four latent variables and gender are considered as the only predictors of log(BMI). The reference categories for each of the four measures are: father or male head in the high social class, family with low financial difficulty, family with low material hardship and parents in stable union, respectively. Estimated coefficients and standard errors from the regression model for log(BMI), using the modal class, 1-step and the 3-step ML approaches are presented in Table 7. It is clear from this analysis that the modal class and 3-step ML approaches yield similar conclusions about the significance of the coefficients, although the estimates differ slightly. This is mainly because each of the four measurement models has a good class separation (entropy>0.7). However, there are some substantial differences between the results from the 1-step and 3-step ML approaches. For the effects of the low level father's social class and high degree of financial difficulty in childhood on log(BMI), the 1-step approach produces estimates that are insignificant and with an opposite sign to those estimated from the 3-step ML approach. This may signal some violation of either the conditional independence or the within-class normality assumptions to which the 1-step approach is more sensitive. Similar to the results from the simulation study, the estimated standard errors of the 1-step approach are slightly higher than those of the 3-step ML approach, while the modal class approach gives the smallest standard errors. This is mainly because latent variables are treated as known in the regression model in the latter approach, and hence the uncertainty in the parameter estimates is underestimated. Based on the 3-step ML results we find that, controlling for gender, children from family backgrounds with fathers in the lower social class and with higher financial difficulty before age 16 tend to have higher and lower values of BMI at age 50, respectively.

DISCUSSION

This paper generalises the 3-step ML approach to estimate the effects of multiple, possibly associated, latent categorical variables on a distal outcome by explicitly specifying a joint distribution of latent variables. The simulation studies show that when all model assumptions are satisfied, the 1-step and 3-step ML approaches perform equally well. When model assumptions are violated, the estimates from both methods are subject to bias, although the 3-step ML approach is less sensitive. The differences in the estimated coefficients in our empirical example are consistent with our findings from the simulation studies. Specifically, when there is within-class non-normality for a continuous Z, skewness of Z is shown to be the worst form of non-normality for both approaches, compared to bimodality and excess kurtosis. Moreover, the results confirmed a major drawback of the 1-step approach as it not only alters the class proportions (shown in Asparouhov & Muthén, 2014a; Bakk & Vermunt, 2016), but also changes the number of classes needed to capture the association among indicators, particularly at low entropy levels. When there is local dependence between Z and the indicators for the latent variables, the 1-step approach leads to greater bias than the 3-step ML approach. This is mainly explained by a tendency to extract too many classes when there is residual correlation between Z and the Us. It should be noted that the extraction of pseudo classes is not necessarily wrong from a theoretical point of view, but one needs to question the validity of such extra classes, which may not be interpretable.

Comparing results of the 3-step ML approach with the BCH approach, in general, we do not observe a consistently better performance of the modified BCH approach in situations where model assumptions are violated, except for the case where the conditional distribution of Z is skewed. If in applications of 3-step ML, a substantial shift in classification from Step 1 to Step 3 is observed, the general 3-step ML approach may not be appropriate and further developments of the BCH approach for more than two latent variables could be helpful in this situation. However, in addition to the severe underestimation of standard errors in the BCH approach (see supplementary material), Bakk and Vermunt (2016) also noted the presence of negative cell frequencies for the BCH approach in an application with a categorical distal outcome and poor class separation. Overall, the development of the 3-step approach is more promising as it is more easily generalised to multilevel models for longitudinal and other forms of clustered data.

Regarding the impact of manipulating design factors (i.e. entropy and sample size) on the amount of bias of the general 3-step ML approach, we find that in cases where distributional assumptions are violated, low entropy levels (when sample size is fixed) and small sample size (when entropy is fixed) lead to poor estimates. In the case where there is local dependence between the distal outcome Z and an indicator U, the performance of the 3-step ML approach is similar at high and low entropy levels for fixed sample size, although a larger sample tends to produce greater bias when entropy is fixed. This could be explained by the fact that class proportions in Step 3 of the 3-step approach are influenced by the inclusion of Z and such influence is more obvious in larger sample sizes.

There are several issues that have not been addressed or discussed in this paper. First, in the three simulation studies and our application, we assume a measurement model where the latent class solution is not influenced by the outcome Z. This is natural when we are interested in a Z that is temporally distal to the indicators U, as is common in longitudinal studies. However, it is possible that Z is an important indicator that helps to identify the latent classes, for example when Z and the Us are measured contemporaneously. In this case, the true data-generating model would take the form of the 1-step model, and we would expect that the 3-step ML approach that excludes Z from the measurement model would lead to incorrect latent class solutions. Second, although simulation results suggest that the 1-step approach tends to extract extra classes when local dependence exists, it should be noted that the approach is also flexible enough to allow for additional pairwise association between Z and the Us without introducing additional classes. In contrast, as Z is only introduced in the last step of the 3-step ML approach, it is less straightforward to adapt this approach to account for local dependence. Third, as we have shown several limitations of the 3-step ML approach when model assumptions do not hold, further research is required that modifies the current approach to improve its robustness. Finally, this paper only considers one distal outcome. Potential extensions are to situations with more than one and possibly mixed types of distal outcomes, or more complex models where external variables can include distal outcomes, mediators, and covariates.

Appendix

Latent GOLD syntax for the empirical example

```
options
        output parameters=effect standarderrors probmeans=posterior profiler
        classification ParameterCovariances frequencies bivariateresiduals iterationdetails;
variables
        dependent m1 nominal 3, m2 nominal 2, m3 nominal 3, m4 nominal 2, logbmi50 continuous;
independent gender nominal coding=2;
        latent l1 nominal 3 coding=3, l2 nominal 2 coding=2,
        l3 nominal 3 coding=3, l4 nominal 2 coding=2;
equations
        l1<-1;l2<-1;l3<-1;l4<-1;
```

```
m1<- (C~wei) 1| l1; m2<- (D~wei) 1| l2; m3<- (E~wei) 1| l3; m4<- (F~wei) 1| l4;
logbmi50<- 1+l1 + l2 + l3 + l4 + gender;
l1<->l2;l1<->l3;l1<->l4; l2<->l3;l2<->l4; l3<->l4;
C={0.824 0.169 0.007
        0.061 0.910 0.028
        0.008 0.093 0.899};
D={0.734 0.266 0.032 0.968};
E={0.854 0.097 0.049
        0.055 0.943 0.002
        0.053 0.003 0.944};
F={0.837 0.163 0.003 0.997};
```

References

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

Asparouhov, T., & Muthén, B. (2014b). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. *Mplus Web Notes*, *21*, 1–22.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: standard errors for correct inference. *Political Analysis*, *22*, 520–540.

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 278–289.

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

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 20–31.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, *42*(4), 757–786.

Bauldry, S., Shanahan, M. J., Russo, R., Roberts, B. W., & Damian, R. (2016). Attractiveness compensates for low status background in the prediction of educational attainment. *PLoS one*, *11*(6), e0155313.

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

Chandola, T., Clarke, P., Morris, J., & Blane, D. (2006). Pathways between education and health: a causal modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(2), 337–359.

Clark, S. L., & Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis.* Retrieved from `www.statmodel.com/download/relatinglca.pdf`

Hobcraft, J. (1998). Intergenerational and life-course transmission of social exclusion: Influences and childhood poverty, family disruption and contact with the police. *CASEpaper 15, Centre for Analysis of Social Exclusion.*

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 1–26.

Muthén, B. (2001). Latent variable mixture modeling. *New Developments and Techniques in Structural Equation Modeling*, 1–33.

Muthén, L., & Muthén, B. (2015). Mplus. *Statistical Analysis with Latent Variables. Version 7.31*, *3*.

Nylund, K. L., Asparouhov, T., & Muthén, B. (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.

Ploubidis, G. B., Silverwood, R. J., DeStavola, B., & Grundy, E. (2015). Life-course partnership status and biomarkers in midlife: Evidence from the 1958 british birth cohort. *American Journal of Public Health*, *105*(8), 1596–1603.

Power, C., & Elliott, J. (2006). Cohort profile: 1958 British birth cohort (national child development study). *International Journal of Epidemiology*, *35*(1), 34–41.

Schoon, I., Sacker, A., & Bartley, M. (2003). Socio-economic adversity and psychosocial adjustment: a developmental-contextual perspective. *Social Science & Medicine*, *57*(6), 1001–1015.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469.

Vermunt, J. K., & Magidson, J. (2013). *LG-Syntax User's Guide: Manual for Latent GOLD 5.0 Syntax Module.*

Table 1

Study 1: Simulation results when all model assumptions are satisfied (N=2000; 500 replications).

Continuous Z		1-step				3-step			
Parameters	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
β_1 (C_1)	2.00	-0.03	0.07	0.07	0.94	-0.01	0.04	0.04	0.95
β_2 (C_2)	-1.50	0.35	0.07	0.07	0.93	0.02	0.01	0.01	0.93
ω_1	0.70	-0.87	0.08	0.06	0.94	0.01	0.07	0.08	0.94
ω_2	0.70	-2.26	0.14	0.08	0.94	-1.05	0.08	0.11	0.81
ω_{12}	0.50	2.28	0.11	0.08	0.93	-1.03	0.09	0.09	0.95
Binary Z		1-step				3-step			
Parameters	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
β_1 (C_1)	1.00	0.49	0.20	0.20	0.96	0.05	0.08	0.08	0.97
β_2 (C_2)	-1.50	0.43	0.19	0.20	0.97	0.10	0.14	0.14	0.94
ω_1	0.70	-0.11	0.07	0.07	0.97	0.01	0.08	0.09	0.94
ω_2	0.70	0.06	0.10	0.11	0.96	0.01	0.09	0.13	0.83
ω_{12}	0.50	0.00	0.08	0.09	0.97	-0.02	0.11	0.11	0.95

Notes: Bias (%)=(Estimate-True)/True × 100%

Table 2

Study 2a: Simulation results for excess kurtosis (N=200, 2000; 500 replications).

		1-step				3-step			
N=200, High entropy	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
β_1 (C_1)	2.00	−4.48	0.30	0.73	0.90	0.60	0.19	0.19	0.94
β_2 (C_2)	−1.50	−7.80	0.30	0.60	0.92	−0.30	0.19	0.19	0.96
N=200, Low entropy									
β_1 (C_1)	2.00	−8.43	0.59	0.77	0.89	−14.38	0.33	0.44	0.84
β_2 (C_2)	−1.50	−11.45	0.59	0.63	0.88	−25.14	0.36	0.79	0.85
N=2000, High entropy									
β_1 (C_1)	2.00	0.03	0.09	0.09	0.95	0.96	0.06	0.06	0.94
β_2 (C_2)	−1.50	−0.09	0.09	0.09	0.97	1.57	0.06	0.06	0.95
N=2000, Low entropy									
β_1 (C_1)	2.00	−52.03	0.14	1.72	0.72	−2.68	0.10	0.10	0.92
β_2 (C_2)	−1.50	55.01	0.15	1.29	0.70	−3.21	0.11	0.11	0.92

Table 3

Study 2b: Simulation results for skewness (N=200, 2000; 500 replications).

		1-step				3-step			
N=200, High entropy	True	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
β_1 (C_1)	2.00	−65.44	0.38	1.64	0.57	−7.31	0.27	0.23	0.90
β_2 (C_2)	−1.50	66.67	0.45	1.44	0.60	−5.66	0.27	0.24	0.94
N=200, Lowentropy									
β_1 (C_1)	2.00	−71.66	1.32	2.15	0.58	−22.73	0.46	0.49	0.56
β_2 (C_2)	−1.50	−24.77	1.34	2.16	0.72	−41.29	0.46	0.49	0.76
N=2000, Highentropy									
β_1 (C_1)	2.00	−84.37	0.13	1.74	0.47	−7.25	0.09	0.07	0.63
β_2 (C_2)	−1.50	72.31	0.13	1.23	0.54	−6.22	0.09	0.07	0.82
N=2000, Lowentropy									
β_1 (C_1)	2.00	−150.11	0.12	0.68	0.03	−28.89	0.14	0.33	0.20
β_2 (C_2)	−1.50	145.00	0.14	0.38	0.03	−38.09	0.14	0.33	0.25

Table 4

Simulation results for the number of classes when there is within-class non-normality of Z. Results are reported in (%) and the selected models are bolded. ssaBIC=sample size adjusted BIC; BLRT=bootstrap likelihood ratio test (average p-values across 500 replications in brackets).

	Entropy	ssaBIC (%)			BLRT (%)		
Scenarios, N=200		3-class	4-class	5-class	3-class (H_0:2-class, H_1:3-class)	4-class (H_0:3-class, H_1:4-class)	5-class (H_0:4-class, H_1:5-class)
Bimodality	High	0	**94**	6	0(0.04)	**100(0.00)**	65(0.30)
	Low	11	**79**	10	40(0.43)	41(0.32)	16(0.58)
Excess kurtosis	High	0	**93**	7	96(0.03)	**100(0.00)**	65(0.29)
	Low	12	**79**	9	41(0.41)	29(0.37)	12(0.60)
Skewness	High	0	**79**	21	98(0.02)	**99(0.01)**	64(0.29)
	Low	4	**76**	20	72(0.20)	63(0.18)	27(0.50)
Scenarios, N=2000							
Bimodality	High	0	**100**	0	100(0.00)	**100(0.00)**	18(0.73)
	Low	0	**84**	16	100(0.00)	**100(0.00)**	30(0.56)
Excess kurtosis	High	0	**98**	2	100(0.00)	**100(0.00)**	31(0.64)
	Low	0	**97**	3	100(0.00)	**100(0.00)**	24(0.60)
Skewness	High	0	48	**52**	100(0.00)	**100(0.00)**	62(0.36)
	Low	0	5	**95**	100(0.00)	100(0.00)	**97(0.03)**

Table 5

Study 3: Estimated coefficients for fitting a 4-class model when there is local dependenc between Z and U_{10}.

	1-step				3-step			
N=500, High entropy	Bias (%)	SE	SD	Coverage	Bias (%)	SE	SD	Coverage
β_3	-11.35	0.32	0.32	0.92	-6.69	0.27	0.30	0.89
β_4	-5.93	0.35	0.35	0.92	-3.54	0.29	0.31	0.93
N=500, Low entropy								
β_3	-59.52	0.59	0.65	0.52	9.01	0.40	0.76	0.70
β_4	-9.70	0.68	0.93	0.66	-6.68	0.44	0.92	0.60
N=2000, High entropy								
β_3	-11.54	0.15	0.16	0.79	-10.98	0.13	0.14	0.72
β_4	-6.04	0.16	0.16	0.82	-6.07	0.14	0.14	0.75
N=2000, Low entropy								
β_3	-60.00	0.30	0.36	0.14	-10.95	0.21	0.48	0.62
β_4	-10.67	0.52	0.70	0.64	-10.64	0.24	0.63	0.50

Table 6

Simulation results for the number of classes when there is local dependence (1-step approach). Results are reported in (%) and selected models are bolded. ssaBIC=sample size adjusted BIC; BLRT=bootstrap likelihood ratio test (average p-values across 500 replications in brackets).

N	Scenario	ssaBIC(%)			BLRT(%)		
High entropy		3-class	4-class	5-class	3-class	4-class	5-class
					(H_0:2-class, H_1:3-class)	(H_0:3-class, H_1:4-class)	(H_0:4-class, H_1:5-class)
500	Independence[a]	0	**89**	13	100(0.00)	100(0.00)	1(0.82)
	Dependence[b]	0	**61**	39	100(0.00)	99(0.00)	14(0.61)
2000	Independence	0	**100**	0	100(0.00)	100(0.00)	2(0.77)
	Dependence	0	2	**98**	100(0.00)	100(0.00)	100(0.00)
Low entropy							
500	Independence	**64**	32	4	68(0.12)	7(0.65)	1(0.82)
	Dependence	31	**52**	17	64(0.13)	26(0.41)	4(0.70)
2000	Independence	12	**88**	0	100(0.00)	**100(0.00)**	2(0.80)
	Dependence	0	18	**82**	100(0.00)	100(0.00)	**97(0.01)**

Notes: a: Independence refers to independence between Z and Us conditional on X;

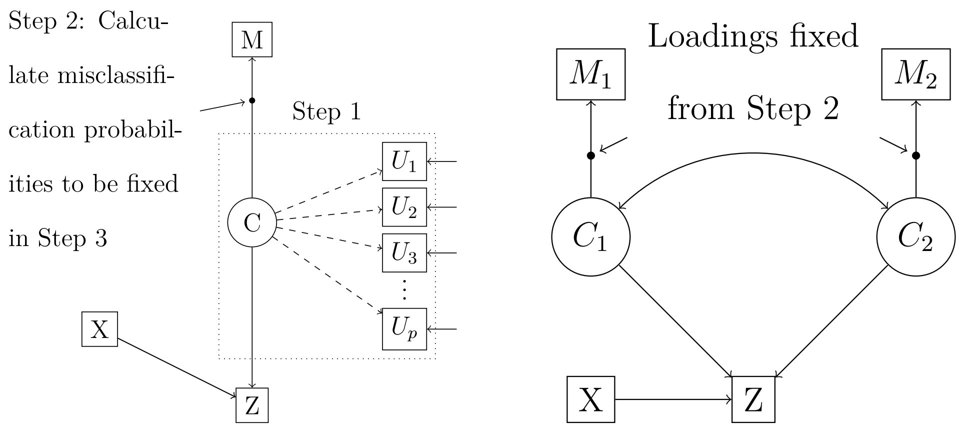
b: Dependence refers to residual correlation between Z and U_{10}.

Table 7

Empirical study: Results for analysis of log(BMI) at age 50: Comparison of the modal class, 1-step and 3-step approaches.

Covariate	Estimate(SE)		
	MC	1-step	3-step
Intercept	3.164**	3.163**	3.163**
	(0.007)	(0.007)	(0.007)
Male	0.208**	0.208**	0.208**
	(0.006)	(0.006)	(0.006)
Latent categorical variables			
Social class of father or male head (ref.=high)			
Low	0.027**	-0.057	0.040**
	(0.010)	(0.046)	(0.018)
Medium	0.030**	0.035**	0.030**
	(0.007)	(0.008)	(0.007))
Financial difficulty (ref.=low)			
High	-0.019**	0.068	-0.037*
	(0.010)	(0.046)	(0.021)
Material hardship (ref.=low)			
Medium	0.008	0.004	0.007
	(0.007)	(0.010)	(0.009)
High	0.008	0.006	0.012
	(0.008)	(0.008)	(0.011)
Family structure (ref.=stable)			
Unstable	0.012	0.003	0.017
	(0.011)	(0.013)	(0.013)

** p<0.05, * p<0.10


Step 2: Calculate misclassification probabilities to be fixed in Step 3
M
Step 1
C
U_1
U_2
U_3
U_p
X
Z
Loadings fixed from Step 2
M_1
M_2
C_1
C_2
X
Z

(a) 1-latent-variable, 3-step ML approach

(b) 2-latent-variable, 3-step ML approach

Figure 1. The 3-step ML approach for 1-latent-variable and 2-latent-variable cases