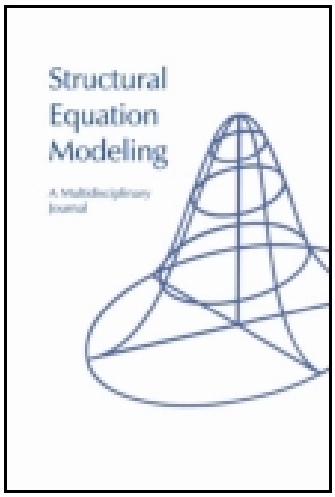

Structural Equation Modeling: A Multidisciplinary Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using Mplus

Tihomir Asparouhov[a] & Bengt Muthén[a]

[a] Mplus

Published online: 09 Jun 2014.

To cite this article: Tihomir Asparouhov & Bengt Muthén (2014) Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using Mplus, Structural Equation Modeling: A Multidisciplinary Journal, 21:3, 329-341, DOI: 10.1080/10705511.2014.915181

To link to this article: http://dx.doi.org/10.1080/10705511.2014.915181

Structural Equation Modeling: A Multidisciplinary Journal, 21: 329–341, 2014

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.915181

Routledge
Taylor & Francis Group

Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using M*plus*

Tihomir Asparouhov and Bengt Muthén
*M*plus

This article discusses alternatives to single-step mixture modeling. A 3-step method for latent class predictor variables is studied in several different settings, including latent class analysis, latent transition analysis, and growth mixture modeling. It is explored under violations of its assumptions such as with direct effects from predictors to latent class indicators. The 3-step method is also considered for distal variables. The Lanza, Tan, and Bray (2013) method for distal variables is studied under several conditions including violations of its assumptions. Standard errors are also developed for the Lanza method because these were not given in Lanza et al. (2013).

Keywords: 3-step estimation, distal outcomes, latent class predictors, mixture modeling, Mplus

In mixture modeling, indicator variables are used to identify an underlying latent categorical variable. In many practical applications we are interested in using the latent categorical variable for further analysis and exploring the relationship between that variable and other, auxiliary observed variables. Two types of analyses are discussed here. The first type of analysis is when we use the observed variable as a predictor of the latent categorical variable. The second type of analysis is using the latent categorical variable as a predictor of an observed variable, which we call a distal outcome. The standard way to conduct such an analysis is to combine the latent class model and the latent class regression model or the distal outcome model into a joint model that can be estimated with the maximum-likelihood estimator. This will be referred to as the one-step method. Such an approach, however, can be flawed because the secondary model could affect the latent class formation and the latent class variable could lose its meaning as the latent variable measured by the indicator variables.

Vermunt (2010) pointed out several disadvantages of the one-step method in the context of predictors (covariates) of the latent class variable:

> However, the one-step approach has certain disadvantages. The first is that it may sometimes be impractical, especially when the number of potential covariates is large, as will typically be the case in a more exploratory study. Each time that a covariate is added or removed not only the prediction model but also the measurement model needs to be reestimated. A second disadvantage is that it introduces additional model building problems, such as whether one should decide about the number of classes in a model with or without covariates. Third, the simultaneous approach does not fit with the logic of most applied researchers, who view introducing covariates as a step that comes after the classification model has been built. Fourth, it assumes that the classification model is built in the same stage of a study as the model used to predict the class membership, which is not necessarily the case. It can even be that the researcher who constructs the typology using an LC model is not the same as the one who uses the typology in a next stage of the study (p. 451).

To avoid such drawbacks, several methods have been developed that can independently evaluate the relationship between the latent class variable and the predictor or distal auxiliary variables. One method is to use the pseudo-class method (see Clark & Muthén, 2009; M*plus* Technical Appendices: Wald Test of Mean Equality for Potential Latent Class Predictors in Mixture Modeling, 2010; Wang, Brown, & Bandeen-Roche, 2005). With this method the latent class model is estimated first, then the latent class variable is multiply imputed from the posterior distribution obtained by

Correspondence should be addressed to Tihomir Asparouhov, Muthén & Muthén, 3463 Stoner Avenue, Los Angeles, CA 90066. E-mail: tihomir@statmodel.com

the latent class analysis (LCA) model estimation. Finally the imputed class variables are analyzed together with the auxiliary variable using the multiple imputation technique developed in Rubin (1987). We call this method the pseudo-class (PC) method. The simulation studies in Clark and Muthén (2009) show that the PC method works well when the entropy is large; that is, the class separation is large. An alternative three-step approach has been developed in Vermunt (2010), expanding ideas presented in Bolck, Croon, and Hagenaars (2004); see also Bakk, Tekle, & Vermunt (2013). This approach is suitable for exploring relationships between the latent class variable and predictor variables. In this approach the latent class model is estimated in a first step using only latent class indicator variables. In the second step, the most likely class variable is created using the latent class posterior distribution obtained during the first step. In the third step the most likely class is regressed on predictor variables, taking into account the misclassification in the second step. Methods are also needed to study the relationship between the latent class variable and distal variables. In a recent paper, Lanza, Tan, and Bray (2013) proposed one such method. Using Bayes theorem, the joint distribution of the latent class variable and the distal variable is represented as a regression of the latent class variable conditional on the distal variable, combined with the marginal distribution of the distal variable.[1]

In this article the three-step method for latent class predictor variables is studied in several different settings including latent transition analysis and is explored under violations of its assumptions such as with direct effects from predictors to latent class indicators. The three-step method is also considered for distal variables. The Lanza et al. (2013) method for distal variables is studied under several conditions including violations of its assumptions. Standard errors are also developed for the Lanza method because these were not given in Lanza et al. (2013). Appendices with M*plus* scripts are referred to in footnotes and are available at www.statmodel.com

The next section presents the Vermunt method for predictors of latent classes and carries out simulation studies of this method. We then extend the method to arbitrary secondary models. The following section presents three-step methods for latent transition analysis. Next, we discuss direct effect violations of the three-step method for predictor variables. Methods for distal outcomes and simulation studies of these methods are discussed next. We present studies of violations of the assumptions for the distal outcome methods before concluding.

PREDICTORS OF LATENT CLASSES

Briefly described, the three-step method for predictors of the latent class variable is as follows. The first step is a regular LCA using only the latent class indicators. In the second step the most likely class variable $\mathcal{N}$, a nominal variable, is created using the latent class posterior distribution obtained during the LCA estimation; that is, for each observation, $\mathcal{N}$ is set to be the class c for which $P(C = c|U)$ is the largest, where U represents the latent class indicators and C is the latent class variable.[2] The classification uncertainty rate for $\mathcal{N}$ is computed as follows:

$$p_{c_1,c_2} = P(C = c_2|\mathcal{N} = c_1) = \frac{1}{N_{c_1}} \sum_{\mathcal{N}_i=c_1} P(C_i = c_2|U_i)$$

where N_{c_1} is the number of observations classified in class c_1 by the most likely class variable $\mathcal{N}$, $\mathcal{N}_i$ is the most likely class variable for the ith observation, C_i is the true latent class variable for the ith observation, and U_i represents the class indicator variables for the ith observation. The probability $P(C_i = c_2|U_i)$ is computed from the estimated LCA model.[3] For example, in the case of a three-class model the probability p_{c_1,c_2} would look like the top part of Table 1, where the p_{c_1,c_2} is in row c_1 and column c_2. We can then compute the probability

$$q_{c_2,c_1} = P(\mathcal{N} = c_1|C = c_2) = \frac{p_{c_1,c_2}N_{c_1}}{\sum_c p_{c,c_2}N_c} \qquad (1)$$

where N_c is the number of observations classified in class c by the most likely class variable $\mathcal{N}$. This shows that $\mathcal{N}$ can be treated as an imperfect measurement of C with measurement error defined by q_{c_1,c_2}.[4] These values are shown in the middle section of Table 1.

[1] All of these methods are obtained in the M*plus* program using the AUXILIARY option of the VARIABLE command. If an auxiliary variable is specified as (R) the PC method will be used and the variable will be treated as a latent class predictor. If an auxiliary variable is specified as (E) the PC method will be used and the variable will be treated as a distal outcome. If an auxiliary variable is specified as (R3STEP) the three-step method will be used and the variable will be treated as a latent class predictor. If an auxiliary variable is specified as (DU3STEP) the three-step method will be used and the variable will be treated as a distal variable with unequal means and variances. If an auxiliary variable is specified as (DE3STEP) the three-step method will be used and the variable will be treated as a distal variable with unequal means and equal variances. If an auxiliary variable is specified as (DCON) or (DCAT), Lanza's method for a continuous or categorical distal variable will be used.

[2] In M*plus* this variable is automatically created using the SAVEDATA command with the option SAVE=CPROB.

[3] In M*plus* the probability p_{c_1,c_2} is automatically computed and can be found in the Results section under the title "Average Latent Class Probabilities for the Most Likely Latent Class Membership (Row) by Latent Class (Column)."

[4] These probabilities are also computed in M*plus* and can be found in the Results section under the title "Classification Probabilities for the Most Likely Latent Class Membership (Row) by Latent Class (Column)."

TABLE 1
Latent Class Probabilities, Classification Probabilities, and Logits for Classification Probabilities

Average Latent Class Probabilities for Most Likely Latent Class Membership (Row) by Latent Class (Column)

	1	*2*	*3*
1	0.839	0.066	0.095
2	0.053	0.845	0.102
3	0.125	0.107	0.768

Classification Probabilities for the Most Likely Latent Class Membership (Row) by Latent Class (Column)

	1	2	3
1	0.830	0.046	0.124
2	0.072	0.811	0.177
3	0.099	0.094	0.807

Logits for the Classification Probabilities for Most Likely Latent Class Membership (Row) by Latent Class (Column)

	1	2	3
1	1.901	−0.990	0.000
2	−0.486	1.936	0.000
3	−2.100	−2.147	0.000

In the third step, the most likely class variable is used as a latent class indicator variable with uncertainty rates prefixed at the probabilities q_{c_1,c_2} obtained in Step 2. That is, the $\mathcal{N}$ variable is specified as a nominal indicator of the latent class variable C with logits $\log\left(q_{c_1,c_2}/q_{K,c_2}\right)$, where K is the last class.[5] These values are shown in the bottom section of Table 1. In this way the measurement error in the most likely class variable $\mathcal{N}$ is taken into account in the third step. In this final step we also include the auxiliary variable. Figure 1 illustrates the third step of the three-step method. The measurement relationships between the latent class variable C and the nominal most likely class variable $\mathcal{N}$ are fixed according to the logit values in the bottom section of Table 1, and the parameters of the multinomial regression of C on the predictor X are estimated.[6]

More details on this approach are available in Vermunt (2010), where it is referred as Modal ML. Here we refer to this approach as the three-step method. In the Vermunt article, this three-step method was used for latent class predictors. In this article we extend the method to distal outcomes, variables that are predicted by the latent class variable.

[5]These logits are also computed in M*plus* and can be found in the Results section under the title "Logits for the Classification Probabilities for the Most Likely Latent Class Membership (Row) by Latent Class (Column)."

[6]Appendix A shows the M*plus* input for the third step and also the input for automatically performing all three steps using the R3STEP option.

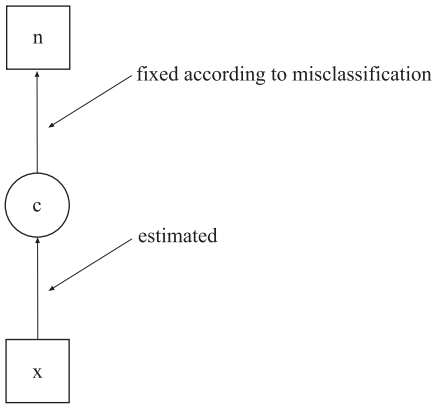

FIGURE 1 Three-step with regression on a predictor.

Simulation Study With a Latent Class Predictor as an Auxiliary Variable

In this simulation study we estimate a two-class model with five binary indicator variables. The distribution for each binary indicator variable U is determined by the usual logit relationship:

$$P\left(U = 1|C\right) = 1/\left(1 + Exp\left(\tau_c\right)\right)$$

where C is the latent class variable that takes values 1 or 2 and the threshold value τ_c is the same for all five binary indicators. In addition we set $\tau_2 = -\tau_1$ for all five indicators. We choose three values for τ_1 to obtain different levels of class separation and entropy. Using the value of $\tau_1 = 1.25$ we obtain an entropy of 0.7, with value $\tau_1 = 1$ we obtain an entropy of 0.6, and with value $\tau_1 = 0.75$ we obtain an entropy of 0.5. The latent class variable is generated with proportions of 43% and 57%. In addition to the preceding latent class model, we also generate a standard normal auxiliary variable as a predictor of the latent class variable through the multinomial logistic regression

$$P\left(C = 1|X\right) = 1/\left(1 + Exp\left(\alpha + \beta X\right)\right)$$

where $\alpha = 0.3$ and $\beta = 0.5$. Five hundred samples of size 500 and 2,000 are generated. The data are analyzed using three different methods: the PC method, the three-step method, and the one-step method.

Table 2 contains the results of the simulation study for the regression coefficient β. The three-step method outperforms the PC method substantially in terms of bias, mean squared error, and confidence interval coverage. The loss of efficiency of the three-step method when compared to the one-step method is minimal. The three-step method also provides good coverage in all cases. The effect of sample size appears to be negligible within the sample size range used in the simulation study. Further simulation studies are needed

TABLE 2
Latent Class Predictor Simulation Study:
Bias/Mean Squared Error/Coverage

N	*Entropy*	*PC*	*Three-Step*	*One-Step*
500	0.7	.13/.023/.84	.01/.015/.95	.01/.014/.95
500	0.6	.20/.044/.59	.00/.019/.96	.01/.017/.96
500	0.5	.28/.083/.24	.02/.029/.95	.03/.028/.97
2,000	0.7	.13/.019/.24	.00/.004/.93	.00/.004/.94
2,000	0.6	.20/.042/.01	.00/.004/.95	.00/.004/.94
2,000	0.5	.29/.085/.00	.01/.007/.94	.01/.006/.95

Note. PC = pseudo-class.

to evaluate the performance for much smaller or much larger sample sizes.[7]

USING M*PLUS* TO CONDUCT THE THREE-STEP METHOD WITH AN ARBITRARY SECONDARY MODEL

In many situations it would be of interest to use the three-step procedure to estimate a more advanced secondary model that includes a latent class variable. In M*plus*, the three-step estimation of the distal outcome model and the latent class predictor model can be obtained automatically using the AUXILIARY option of the VARIABLE command as illustrated earlier. However, for more advanced models, the three-step procedure has to be implemented manually, meaning that each of the three steps is performed separately. In this section we illustrate this manual three-step estimation procedure with a simple auxiliary model where the latent class variable is a moderator for a linear regression. The joint model, which combines the measurement and the auxiliary models, is visually presented in Figure 2.

Suppose Y is a dependent variable and X is a predictor and suppose that a three-category latent class variable C is measured by 10 binary indicator variables. We want to estimate the secondary model independently of the latent class measurement model part. The secondary model is described as follows

$$Y = \alpha_c + \beta_c X + \varepsilon$$

where both coefficients α_c and β_c depend on the latent class variable C. The measurement part of the model is a standard LCA model described by

$$P\left(U_p = 1\,|C\right) = 1/\left(1 + Exp\left(\tau_{cp}\right)\right)$$

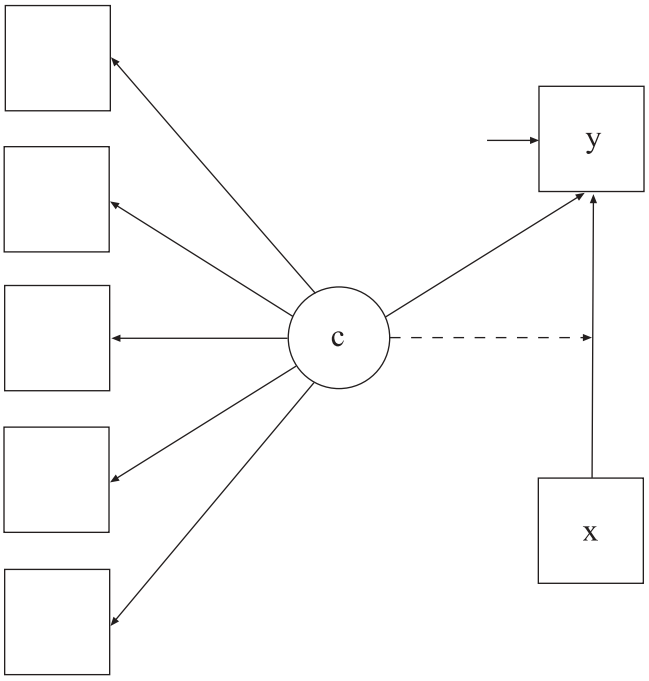

FIGURE 2 Linear regression auxiliary model.

for $p = 1, \ldots, 10$ and $c = 1, \ldots, 3$. We generate a sample of size 1,000 using equal classes and the following parameter values $\tau_{1p} = -1$, $\tau_{2p} = 1$, $\tau_{3p} = 1$ for $p = 1, \ldots, 5$, $\tau_{3p} = -1$ for $p = 6, \ldots, 10$. The parameters in the secondary model used for generating the data are as follows: X and ε are generated as standard normal and the linear model parameters are as follows $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = -1$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, $\beta_2 = 0$.[8]

The first step in the three-step estimation procedure is to estimate the measurement part of the joint model; that is, the latent class model. Thus in Step 1 we estimate the LCA model with the 10 binary indicator variables and without the secondary model.[9]

In Step 2 of the estimation we have to determine the measurement error for the most likely class variable N. This measurement error will be used in the last step of the estimation. In the Step 1 output file we find the following 3 $\times$ 3 table titled Logits for the Classification Probabilities the Most Likely Latent Class Membership (Row) by Latent Class (Column); see the bottom part of Table 1. This table contains $\log\left(q_{i,c}/q_{3,c}\right)$, where the probabilities q_{c_1,c_2} are computed using Equation 1.

The final step in the three-step estimation procedure is estimating the desired auxiliary model where the latent class variable is measured by the most likely class variable N and the measurement error is fixed and prespecified to the values

[7]Appendix B contains an input file for conducting a simulation study with a latent class predictor auxiliary variable.

[8]Appendix C contains the input file for generating this data set.

[9]The input file for this estimation is given in Appendix D along with an explanation of the commands.

TABLE 3
Final Estimates From the Manual Three-Step Estimation With Linear Regression Auxiliary Model

Parameter	*True Value*	*Estimate*	*Standard Error*
α_1	0	0.022	0.068
β_1	0.5	0.490	0.067
α_2	1	1.083	0.072
β_2	−0.5	−0.452	0.063
α_3	−1	−1.078	0.070
β_3	0	0.092	0.059

computed in Step 2.[10] The estimates obtained in this final stage are presented in Table 3. These estimates are very close to the true parameter values and we conclude that the three-step procedure works well for this example. This example also illustrates how M*plus* can be used to estimate an arbitrary auxiliary model with a latent class variable in a three-step procedure where the measurement model for the latent class variable is estimated independently of the auxiliary model.

ESTIMATING LATENT TRANSITION ANALYSIS USING THE THREE-STEP METHOD

In latent transition analysis (LTA), several latent class variables are measured at different time points and the relationship between these variables is estimated through a logistic regression. A three-step estimation procedure can be conducted for the LTA model where the latent class variables are estimated independently of each other and are formed purely based on the latent class indicators at the particular point in time. This estimation approach is desirable in the LTA context because the one-step approach has the drawback that an observed measurement at one point in time affects the definition of the latent class variable at another point in time. We illustrate the estimation with two different examples. The first example is a simple LTA model with two latent class variables. The second example is an LTA model with covariates and measurement invariance. To achieve measurement invariance an additional step is required, so we illustrate this separately. Note, however, that both examples can easily accommodate covariates. Thus to estimate an LTA model with covariates but without measurement invariance the first approach should be used because it is simpler.

Simple LTA

For illustration purposes we consider an example with two latent class variables C_1 and C_2 each measured by five binary indicators. The coefficient of interest, estimated in the three-step approach, is the regression coefficient of C_2 on C_1.

As a first step, an LCA model is applied to the latent class indicators of time point 1 and another LCA model is applied to the latent class indicators of time point 2. These two analyses generate the most likely class variables N_1 and N_2. These variables are the observed variables in the last step that uses log ratios as in the preceding section to take into account classification errors. The last step estimates the C_1, C_2 parameters.[11]

The three-step approach produces an estimate of 0.645 for the regression of C_2 on C_1 with a standard error of 0.175 where the true value is 0.5; that is, the estimate is close to the true value. Simulation studies are currently not easy to conduct in M*plus* using the manual approach because the log ratios need to be computed for every replication. A small simulation study conducted manually using 10 replications revealed that the average estimate across the 10 replications is 0.486, the coverage was 100%, and the ratio between the average standard errors and standard deviation is 1.18. Thus we conclude that the three-step estimator performs well for the LTA model. This approach can also be used for three-step LTA estimation with more than two latent class variables and also with covariates that will be used only in the third step.

LTA With Covariates and Measurement Invariance

In addition, it is possible to estimate the LCA measurement model under the assumption of measurement invariance, which implies that the threshold parameters are equal across time. The approach illustrated in the previous section is inadequate and cannot be used to estimate the three-step LCA with measurement invariance because the LCA at the different time points is estimated in different input files. It is possible, however, to estimate three-step LTA with measurement invariance and also include a covariate X.

In a first step, the two LCA models at the two time points are estimated in one analysis but independently of each other while holding all thresholds equal to obtain the LTA model with measurement invariance. Even though we are interested in an auxiliary model estimation where C_2 is regressed on C_1, at this point of the estimation we estimate the model without such a regression in line of the three-step methodology. The actual regression of C_2 on C_1 will be estimated in the last step of the three-step estimation. Thus in this step we estimate a model assuming that C_1 and C_2 are independent. Note that if the measurement invariance is removed from this model, the estimation of C_1 and C_2 measurement models would be identical to the one from the previous section, where C_1 and C_2 measurement models are estimated independently of each other and in two separate files. This is because without the measurement invariance the log-likelihood of the joint model will be split in two independent parts that can be estimated separately.

[10]The input file for our example is provided in Appendix E along with explanations of the commands.

[11]Appendices F, G, H, and I illustrate how the entire process is carried out in M*plus*.

In the second step, the LCA estimation is done for each set of latent class indicators separately to obtain the most likely class variables and their classification errors. In these two LCA models the measurement parameters are held fixed at the values estimated in the first step. We manually calculate the log ratios from the error tables for N_1 and N_2 as was done in the preceding section.

In the third and final step N_1 and N_2 are used as C_1 and C_2 indicators with parameters fixed at the second step log ratios. This input now contains the auxiliary model that contains the regression of C_2 on C_1 as well as the regression of C_1 and C_2 on X.[12]

In this particular example the true value for C_1 on C_2 is 0.5 and the three-step estimate for that parameter is 0.63(0.19). The true value for C_2 on X is -0.5 and the three-step estimate is $-0.58(0.07)$. The true value for C_2 on X is 0.3 and the three-step estimate is 0.22(0.08). All parameters of the auxiliary model are covered by the confidence intervals obtained by the three-step estimation procedure and thus we conclude that the three-step method works well for the LTA model with measurement invariance.

SIMULATION STUDIES OF OMITTED DIRECT EFFECTS FROM THE LATENT CLASS PREDICTOR AUXILIARY VARIABLE

In this section we study the ability of the three-step method to absorb misspecifications in the measurement model due to omitted direct effects from a covariate. Vermunt (2010) suggested that the three-step estimation might be a more robust estimation method in that context. We consider two different situations: direct effects in LCA and direct effects in growth mixture models (GMMs).

Direct Effects in LCA

Data are generated with 10 binary indicators using the following equations:

$$P(C = 1|X) = 1/(1 + Exp(\alpha + \beta X))$$

$$P(U_p = 1|C) = 1/(1 + Exp(\tau_{pc} + \gamma_{pc} X)).$$

The second equation shows that there are direct effects from X to the indicator variables. For data generation purposes, almost all of the parameters γ_{pc} are zero. To vary the magnitude of direct effect influence we vary the number of nonzero direct effects. All nonzero direct effects γ_{pc} are set to 1. We generate different samples with L direct effects for $L = 1, 2, \ldots, 5$. All nonzero direct effects are in Class 1. To obtain different entropy values we use $\tau_{pc} = \pm 1.25$, which leads to entropy of 0.9 and $\tau_{pc} = \pm 0.75$, which leads to entropy of 0.6. The values of α and β are as in the previous section. We generate samples of size 2,000.

The generated data are analyzed with three different methods. Method 1 ignores the direct effect in the LCA measurement model and analyzes the regression of C on X using the three-step procedure. Method 2 includes the direct effect in the LCA measurement model and analyzes the regression of C on X using the three-step procedure. Method 3 is the one-step approach that includes the direct effects and estimates the regression of C on X together with the measurement model in one joint model.

Table 4 contains the bias and coverage simulation results for the regression parameter β. It is clear from these results that the ability of the three-step approach to estimate the correct relationship between C and X is somewhat limited. Method 1, which ignores the direct effects and estimates the β coefficient with the three-step approach, performs quite poorly when the number of direct effects is substantial but it has good performance when the number of direct effects is small and the entropy is large. Using this method has the fundamental flaw that the latent variable C cannot be measured correctly if the covariate X is not included in the model. This is because there is a violation in the identification condition for the latent class variable that postulates that the measurement indicators are independent given C. The indicator variables are actually correlated beyond the effect of C through the direct effects from X. Therefore, if there is a sufficient number of omitted direct effects, the latent class variable cannot be measured well only by the indicator variables. That, in turn, leads to substantial biases in the C on X regression using the three-step approach. More extensive discussion on the effects of omitted direct effects in the growth mixture context can be found in Muthén (2004).

Method 2, which uses a properly specified measurement model that includes the direct effects, performs much better, but biases are found with this three-step method as well

TABLE 4
Latent Class Analysis With Direct Effects: Absolute Bias and Coverage

No. of Direct Effects	*Entropy*	*Method 1 Three-Step Excluding Direct Effects*	*Method 2 Three-Step Including Direct Effects*	*Method 3 One-Step*
1	0.9	0.02 (.92)	0.02 (.94)	0.01 (.94)
2	0.9	0.04 (.88)	0.00 (.94)	0.01 (.94)
3	0.9	0.08 (.68)	0.01 (.96)	0.01 (.94)
4	0.9	0.15 (.24)	0.01 (.97)	0.01 (.95)
5	0.9	0.25 (.04)	0.00 (.94)	0.01 (.95)
1	0.6	0.08 (.79)	0.05 (.83)	0.01 (.95)
2	0.6	0.19 (.30)	0.04 (.92)	0.01 (.97)
3	0.6	0.38 (.00)	0.01 (.92)	0.01 (.97)
4	0.6	0.56 (.00)	0.07 (.81)	0.01 (.99)
5	0.6	0.76 (.00)	0.08 (.80)	0.01 (.97)

[12] Appendices J through N show how to generate data and carry out the analysis steps.

when the entropy is 0.6. In contrast, the three-step procedure performed very well at that entropy level when direct effects were not present. Method 2 can also suffer from incorrect classification, but to a much smaller extent than Method 1. In this situation, even with all direct effects included, the effect of X on U is not captured completely because the measurement model does not include the effect of X on C, which will have to be absorbed by the direct effects. That might lead to misestimation of some of the parameters, which in turn will lead to biases in the formation of the latent classes and biases in the auxiliary model estimation.[13]

The one-step approach performs well in all cases. The analyses indicate that the three-step approach has a limited ability to deal with direct effects and thus when substantial direct effects are found, those effects should be included in the measurement model for the latent class variable even with the three-step approach. In the preceding simulation study the direct effects are quite large and in many practical applications the direct effect could be much smaller. Further exploration is necessary to evaluate the performance of the three-step methods for various levels of direct effect.

Direct Effects in Growth Mixture Models

The impact of direct effects on the three-step estimation can also be seen in the context of GMMs (Muthén, 2004; Muthén & Asparouhov, 2009; Muthén & Shedden, 1999) when the direct effect is not on the observed variables but it is on the growth factors. Consider the following GMM:

$$Y_t = I + S \cdot t + \varepsilon_t$$

where Y_t are the observed variables and I and S are the growth factors that also identify the latent class variable C through the following model:

$$I\,|C = \alpha_{1c} + \beta_{1c}X + \xi_1$$

$$S|C = \alpha_{2c} + \beta_{2c}X + \xi_2$$

where X is an observed covariate. This model simply postulates that the latent classes are determined by the pattern of growth trajectory; that is, the latent class variable determines the mean of the intercept and the slope growth factors, but individual variation is allowed. This GMM is essentially the measurement model for the latent class variable C. In this situation we are again interested in estimating with the three-step approach the relationship between C and X independently of the measurement model; that is, we want to estimate the logistic regression model

$$P\,(C = 1|X) = 1/\,(1 + Exp\,(\alpha + \beta X))\,.$$

We generated 100 samples of size 5,000 using the following parameter values: $\alpha = 0$, $\beta = 0.5$, $Var\,(\varepsilon_t) = 1$, $Var\,(I) = 1$, $Var\,(S) = 0.4$, $Cov\,(I, S) = 0.2$, $\alpha_{21} = 1$, $\alpha_{22} = -0.5$, and $t = 0, 1, \ldots, 4$. We also vary the values of α_{1c} to obtain different entropy levels. Choosing $\alpha_{11} = 1$, $\alpha_{12} = -1$ yields entropy of 0.6. Choosing $\alpha_{11} = 2$, $\alpha_{12} = -2$ yields entropy of 0.85. Choosing $\alpha_{11} = 3$, $\alpha_{12} = -3$ yields entropy of 0.95. We also want to explore different types of direct effects so we generate three different types of data. Type 1 uses no direct effects; that is, $\beta_{1c} = \beta_{2c} = 0$. Type 2 uses the same direct effects across the two classes $\beta_{1c} = 1$ and $\beta_{2c} = 0.2$; that is, the direct effect is independent of the latent class variable. Type 3 uses different direct effects across the two classes $\beta_{11} = 1$, $\beta_{21} = 0.2$, and $\beta_{12} = \beta_{22} = 0$. As in the LCA simulation study, we use different estimation methods. Method 1 is a three-step method that uses only the growth model as the measurement model, Method 2 uses the growth model as the measurement model but includes the direct effects from X to the growth factors. Method 3 is the one-step approach using the direct effects and the regression from C on X.

The results for the β estimates are presented in Table 5. Again we see here that Method 1 works well but only if there are no direct effects from X on the measurement model (Type 1 data). The biases for Type 2 and 3 decrease substantially when the the entropy increases but these biases are too high even with entropy of 0.85. Method 2 performed much better than Method 1; thus including covariates in the measurement model is important here as well, but the biases are unacceptable when the entropy is 0.6. Method 2 seems to perform better for Type 2 data where the direct effects are independent of C, even though the direct effects are bigger. Method 3, as expected, performed well. This method uses the maximum likelihood estimator for the correctly specified model.

The identification of the latent class variable is more complicated in the GMM than in the LCA model. The local independence assumption of the LCA model is not present in the GMM. Nevertheless we see the same pattern: If the covariates have direct effects on the measurement model, these effects should be included for the three-step approach to work well. More simulation studies are needed to evaluate

TABLE 5
Growth Mixture Model With Direct Effects: Absolute Bias and Coverage

Entropy	*Method 1 Type 1*	*Method 1 Type 2*	*Method 1 Type 3*	*Method 2 Type 2*	*Method 2 Type 3*	*Method 3 Type 3*
0.6	0.00 (.97)	0.68 (.00)	0.49 (.00)	0.18 (.00)	0.24 (.00)	0.00 (.93)
0.85	0.04 (.95)	0.35 (.00)	0.23 (.00)	0.02 (.92)	0.09 (.26)	0.00 (.96)
0.95	0.00 (.95)	0.12 (.06)	0.07 (.32)	0.00 (.95)	0.01 (.90)	0.00 (.94)

[13]The use of M*plus* to carry out this approach is illustrated in Appendix O.

the impact of the size of the direct effects on the three-step estimation.

DISTAL OUTCOME AUXILIARY VARIABLE

Turning to the case of a distal outcome, two approaches are studied: the three-step method applied to distal outcomes and Lanza's method. The three-step method for a distal outcome uses the approach presented earlier. For example, Figure 2 without the X variable is exactly the situation considered here.[14]

A new method for the estimation of auxiliary distal outcomes has been proposed in Lanza et al. (2013). This method has the advantage over the three-step method that it does not allow for the distal outcome to dramatically change the class membership for individual observations. The method can be used with a categorical or a continuous distal outcome. The idea behind the method is that after the LCA model is estimated we can estimate an auxiliary model where the distal outcome X is used as a latent class predictor within a multinomial logistic regression in addition to the the original measurement LCA model. The auxiliary model is used to obtain the conditional distribution $P(C|X)$ as well as the marginal distribution $P(C)$. Using also the sample distribution of X one can easily derive the desired conditional distribution $P(X|C)$ by applying Bayes's theorem:

$$P(X|C) = \frac{P(X)\,P(C|X)}{P(C)}. \tag{2}$$

If X is a continuous variable the mean parameters can then be estimated within each class and if it is a categorical variable the probabilities for each category can be estimated within each class.

Lanza's method has a number of limitations. The method can only be used with distal auxiliary variables. In addition, the method cannot have a latent class measurement model that already includes latent class predictors. The original article by Lanza et al. (2013) does not include standard error computations. Such standard errors are easy to obtain if the auxiliary variable is categorical using the delta method in Equation 2, but in the continuous case it is not very clear how to compute the standard errors because is the sample distribution. As implemented in M*plus*, Lanza's method uses approximate standard errors for continuous distal outcomes by estimating the mean and variance within each group as well as the within-class sample size. Standard errors are then computed as if the mean estimate is the sample mean. For both continuous and categorical distal outcomes, M*plus* computes an overall test of association using Wald's test as well as pairwise class comparisons between the auxiliary variable means and probabilities. There is a slight difference between the continuous distal outcome estimation described in Lanza et al. and the method implemented in M*plus*. Lanza's method uses kernel density estimation to approximate the density function for the distal outcome, whereas the method implemented in M*plus* uses the sample distribution for the auxiliary variable directly. The two methods, however, should yield similar results.[15]

Simulation Study With a Continuous Distal Auxiliary Outcome: Comparing the three-Step and Lanza Methods

In this simulation study we estimate a two-class model with five binary indicator variables. The distribution for each binary indicator variable U is determined by the usual logit relationship

$$P(U = 1\,|C) = 1\big/(1 + Exp(\tau_c))$$

where C is the latent class variable that takes values 1 or 2 and the threshold value is the same for all five binary indicators. In addition we set $T_2 = -T_1$ for all five indicators. We choose three values for to T_1 obtain different levels of class separation and entropy. Using the value of $T_1 = 1.25$ we obtain an entropy of 0.7, with value we obtain an entropy of 0.6, and with value $T_1 = 0.75$ we obtain an entropy of 0.5. The latent class variable is generated with proportions of 43% and 57%. In addition to the preceding latent class model, we also generate a normally distributed distal auxiliary variable with mean 0 in Class 1 and mean 0.7 in Class 2 and variance 1 in both classes. We apply the PC method, the three-step method, Lanza's method, and the one-step method to estimate the mean of the auxiliary variable in the two classes.

Table 6 presents the results for the mean of the auxiliary variable in Class 2. We generate 500 samples of size 500 and 2,000 and analyze the data with the four methods. It is clear

[14] In M*plus* this can be analyzed using the "manual" method shown earlier or all three steps carried out automatically using either of two auxiliary options. If an auxiliary variable is specified as (DU3STEP) the three-step method will be used and the variable will be treated as a distal outcome with unequal means and variances. If an auxiliary variable is specified as (DE3STEP) the three-step method will be used and the variable will be treated as a distal outcome with unequal means and equal variances. The equal variance estimation is useful for situations when there are small classes and the distal outcome estimation with unequal variance might have convergence problems due to near-zero variance within class. For example, if the distal outcome is binary this can occur quite easily. However the equal variance option should not be used in general because it could lead to biases in the estimates and the standard error if the equal variance assumption is violated.

[15] If an auxiliary variable is specified as (DCON) the Lanza et al. (2013) method will be used and the variable will be treated as a distal continuous outcome. If an auxiliary variable is specified as (DCAT) the Lanza et al. method will be used and the variable will be treated as a distal categorical outcome.

TABLE 6
Distal Outcome Simulation Study:
Bias/Mean Squared Error/Coverage

N	*Entropy*	*PC*	*Three-Step*	*Lanza*	*One-step*
500	0.7	.10/.015/.76	.00/.007/.95	.00/.006/.92	.00/.006/.94
500	0.6	.16/.029/.50	.01/.008/.94	.00/.007/.89	.00/.007/.94
500	0.5	.22/.056/.24	.03/.017/.86	.00/.012/.80	.01/.012/.96
2,000	0.7	.10/.011/.23	.00/.002/.93	.00/.002/.89	.00/.002/.93
2,000	0.6	.15/.025/.03	.00/.002/.93	.00/.002/.87	.00/.002/.94
2,000	0.5	.22/.051/.00	.00/.004/.91	.00/.003/.80	.00/.003/.94

Note. PC = pseudo-class.

from the results in Table 6 that the three-step procedure outperforms the PC procedure substantially in terms of bias, mean squared error, and confidence interval coverage. When the three-step procedure is compared to the one-step procedure, it appears that the loss of efficiency is not substantial especially when the class separation is good (entropy of 0.6 or higher). The loss of efficiency can be seen, however, in the case when the entropy is 0.5 and the sample size is 500. The three-step procedure also provides good confidence interval coverage. Lanza's method appears to be slightly better than the three-step method in terms of bias and mean square error, but in terms of coverage the three-step method appears to be better. The effect of the sample size appears to be negligible in the sample size range from 500 to 2,000. Further simulation studies are needed to evaluate the performance of the three-step procedure and Lanza's method for much smaller or much larger sample sizes.[16]

Next we conduct a simulation study to compare the performance of the two different three-step approaches. The two approaches differ in the third step. The first approach estimates different means and variance for the distal variable in the different classes, and the second approach estimates different means but equal variances. The second approach is more robust and more likely to converge but might suffer from the misspecification that the variances are equal in the different classes. We use the same simulation as earlier except that we generate a distal outcome in the second class with variance 20 instead of 1. The results for the mean in the second class are presented in Table 7. It is clear from these results that the unequal variance three-step approach is superior, particularly when the class separation is poor (entropy level of 0.6 or less). The equal variance approach can lead to severely biased estimates when the class separation is poor and the variances are different across classes. The results obtained in this simulation study might not apply if the ratio between the variances is much smaller. Further simulation studies are needed to determine exactly what level of discrepancy between the variances leads to an accuracy advantage for the unequal variance three-step approach.

[16]Appendix P contains an M*plus* input file for conducting a simulation study with a distal auxiliary variable.

TABLE 7
Distal Outcome Simulation Study Comparing Equal and Unequal Variance Three-Step Methods: Bias/Mean Squared Error/Coverage

N	*Entropy*	*Three-Step Equal Variance*	*Three-Step Different Variance*
500	0.7	.05/.147/.95	.00/.099/.94
500	0.6	.06/.174/.96	.00/.099/.95
500	0.5	.12/.822/.93	.01/.101/.95
2,000	0.7	.05/.040/.92	.00/.027/.92
2,000	0.6	.09/.056/.92	.00/.027/.93
2,000	0.5	.11/.094/.95	.00/.029/.92

Simulation Study With distal Categorical Outcome Using Lanza's Method

In this section we conduct a simulation study to evaluate the performance of Lanza's method with categorical auxiliary outcome. We generate data from a three-class LCA model where the latent class variable is measured by 10 binary variables. In Class 1, $P(U_i = 0|C = 1) = 1/(1 + Exp(\tau))$ for all indicator variables U_i. In Class 3, $P(U_i = 0|C = 3) = 1/(1 + Exp(-\tau))$ for all indicator variables U_i. In Class 2, $P(U_i = 0|C = 2) = 1/(1 + Exp(\tau))$ for $i = 1, \ldots, 5$ and $P(U_i = 0|C = 2) = 1/(1 + Exp(-\tau))$ for $i = 6, \ldots, 10$. We use two τ values in the generation process. If $\tau = 0.75$ the entropy for the LCA model is $I = 0.5$ and if $\tau = 1$ the entropy for the LCA model is 0.65. The class probabilities $P(C = j)$ for the three classes are 0.32, 0.44, and 0.24, respectively. The auxiliary variable X is a categorical variable with 6 categories. The probabilities $P(X = k|C = 1)$ of these six categories in Class 1 are 0.18, 0.09, 0.23, 0.23, 0.09, and 0.18. In Class 2 these probabilities are 0.08, 0.65, 0.05, 0.03, 0.11, and 0.08. In Class 3 these probabilities are 0.07, 0.11, 0.32, 0.23, 0.09, and 0.18. We generate 100 samples of size $N = 200$, $N = 500$, and $N = 2{,}000$ using both entropy levels and analyze the X variable as an auxiliary variable.

The results of the simulation are presented in Table 8. We present the bias and the coverage for Category 2 in the three different classes $p_{2j} = P(X = 2|C = j)$.

The results for the remaining categories are similar. The simulation shows that the estimates are unbiased and the coverage is near the nominal level of 95%. Some small sample

TABLE 8
Categorical Distal Outcome Simulation Study Using Lanza's Method: Absolute Bias(Coverage)

N	*Entropy*	p_{21}	p_{22}	p_{23}
200	0.5	.03 (.62)	.08 (.74)	.07 (.74)
200	0.65	.00 (.87)	.01 (.90)	.02 (.90)
500	0.5	.01 (.87)	.00 (.99)	.01 (.91)
500	0.65	.01 (.88)	.00 (.93)	.00 (.92)
2,000	0.5	.00 (.91)	.00 (.95)	.01 (.95)
2,000	0.65	.00 (.93)	.00 (.95)	.00 (.95)

size bias can be seen here when $N = 200$, in particular when the entropy is low at $I = 0.5$. These biases are, however, mostly due to the quality of the estimation of the measurement model, which also has estimation biases when the sample and the entropy are small. Overall we conclude that Lanza's method as well as the standard error method based on the delta method implemented in M*plus* work well. Note here that in the categorical distal outcome case, Lanza's method does not rely on the multinomial model assumption. In this case, the multinomial model together with the marginal distribution model for the auxiliary variable yield a saturated bivariate model for the two categorical variables. The estimated joint distribution model for the latent class and the distal outcome variables is the full contingency unrestricted model. Thus there are no underlying assumptions in this case, as is the case in the continuous distal outcome situation.

DISTAL OUTCOME ESTIMATION FAILURES

In this section we discuss different situations in which the distal outcome estimation methods fail. We first present a simulated example where the one-step and the three-step methods fail due to change in the latent class variable when the auxiliary variable is added to the latent class measurement model. We then present a simulated example where Lanza's method fails due to an incorrect multinomial model assumption.

Failure Due to Change in the Latent Class Variable

In this section we describe a distal outcome simulated example that illustrates the potential failure when using the one-step and the three-step methods. In this example Lanza et al. (2013) does not fail. This shows that the Lanza et al. method might be more robust in practical situations.

We generate a data set of size $N = 5{,}000$ according to a two-class LCA model with five binary indicators $U_i, i = 1, \ldots 5$ using $P(U_i|C = 1) = 0.73$ and $P(U_i|C = 2) = 0.27$. The two latent classes are equally likely $P(C = j) = 0.5$, for $j = 1.2$. To that data set we add a continuous variable X that has a bimodal distribution $0.75 \cdot N(0, 0.01) + 0.25 \cdot N(1, 0.01)$; that is, the bimodal distribution is a mixture of two normal distributions with means 0 and 1 and variance 0.01 and with weights 0.75 and 0.25. The continuous variable X is generated as an independent variable. The variable is independent of the class indicators U_i as well as the latent class variable C. Thus if we analyze the variable X as an auxiliary distal outcome variable, we expect to see no significant effect from C to X; that is, if $m_j = E(X|C = j)$ is the class-specific mean of X, we expect the mean difference parameter $m = m_2 - m_1$ to be statistically insignificant from 0. In addition we expect the latent class proportions $P(C = 1)/P(C = 2)$ to be near 1.

The results of this analysis are presented in Table 9. We analyze the simulated data with the four different methods available in M*plus*, one-step, three-step with unequal variances, the PC method, and the Lanza et al. (2013) method. In addition we analyze the model with the three-step manual procedure described previously. Both the one-step procedure and the three-step manual procedures failed.

The class allocation changed from equal classes to a ratio of 3, which corresponds to the bimodal distribution weights suggesting that the latent class variable has changed its meaning and is now used to fit the bimodal distribution of the auxiliary variable and the original measurement model is ignored. This happens because the methods use the maximum-likelihood estimation. Ultimately the log-likelihood will be maximized and in this particular example the log-likelihood benefits more by fitting the distal outcome variable rather than the measurement model. Most important, the one-step and the three-step manual procedures failed in the distal outcome estimation. Both methods find large and statistically significant effects from the latent class variable on the auxiliary distal outcome where such an effect does not exist, according to how the data were generated. This effect was found because the latent class variable meaning changed.

Interestingly, the M*plus* automated three-step procedure did not fail. The difference between the automated and the manual procedure is in the starting values. The manual procedure will use a number of random starting values and by default M*plus* will use 20 to guarantee that the global maximum is found. On the other hand the automated procedure will not use random starting values and instead will use as starting values only the parameters obtained in the first-step estimation when the latent class measurement model is estimated separately without the auxiliary variable. Using such starting values, it is very likely that a local optimum will be reached that preserves the meaning of the latent class variable from the first step if such a local optimum exists. If that local optimum is also a global optimum, the manual three-step procedure and the automated three-step procedure will yield the same result; however, if the local optimum is not a global optimum, the two procedures will yield different results. In our simulated data set the local optimum is not a global optimum. The log-likelihood obtained

TABLE 9
Distal Outcome Simulated Example

Method	*m*	*p value*	$P(C = 1)/P(C = 2)$
One-step	0.986	0.007	2.8
Three-step manual	0.986	0.007	2.8
Three-step	0.013	0.858	1.0
PC	0.019	0.492	1.0
Lanza	0.019	0.492	1.0

Note. PC = pseudo-class.

with the manual three-step procedure is −770.197 and it is much better than local optimum obtained with the automated procedure of – 1300.201.

There are two issues that we need to address related to local and global optima. First one might ask if it is a good statistical practice to use the local optimum instead of the global optimum. Obviously in this particular example it makes sense, because, the local optimum yields unbiased estimates for the distal outcome model whereas the global optimum does not. The fact is that it is also a theoretically solid approach. Using a local optimum instead of a global optimum usually is equivalent to adding parameter constraints to the model. In our example we could have added to the model estimation the constraint that the two classes' probabilities are between 45% and 55%. Given that the LCA class without the auxiliary variable yields two almost equal classes such a parameter constraints seems reasonable. If the parameter constraints are added, then the global optimum is unacceptable and the local optimum becomes the global optimum and therefore an acceptable solution. In fact, what we obtained in this example as the global optimum is not really the global optimum. Given that the variance of the distal outcome is unconstrained, a class allocation where one of the classes has a single observation and a variance of 0 has a likelihood of infinity; that is, the log-likelihood does not have a global maximum in a completely unconstrained sense.

The second issue we have to address is the fact that a local optimum corresponding to the original latent class model might not exist. This actually is very likely to happen when the number of classes is large and larger than what is supported by the data; that is, when the classes are poorly identified and the entropy of the Step 1 latent class model is low, and thus the nominal indicator S is a weak class indicator. In that case the three-step method simply fails.[17]

The Table 9 results also show that the PC method and the Lanza method are more robust estimation methods than the one-step and the three-step methods. Because these methods do not include new dependent variables in the final model estimation, they are less likely to alter the meaning of the latent class variable. Both methods yield the correct result that the effect of the latent class on the auxiliary variable is not statistically significant.

[17] A simple check is implemented in M*plus* to verify that this failure does not occur and if it does the method will not report any results because those results are likely to be incorrect, similar to the results reported in Table 9. This consistency check is computed as follows. Each observation is classified into the most likely class using both the first step model and the third step model. If more than 20% of the observations in the Step 1 class move to a different class in Step 3, then the three-step estimation is determined to be inconsistent and no results are reported. Because this check is already implemented in M*plus* version 7.1, it is safe to use the automatic three-step procedure without investigating further the class formation.

Failure Due to Incorrect Multinomial Model Assumptions

Lanza's method is based on the underlying assumption that we can estimate the joint distribution of the latent class variable and the auxiliary variable through estimating a multinomial regression model where the latent class variable is regressed on the auxiliary variable. This multinomial model, however, might not hold. In that case, the estimated class-specific means for the auxiliary distal variable might be biased. Note that in the earlier simulation studies the multinomial model does not hold. Nevertheless we obtained unbiased results. Apparently, the multinomial model is quite robust in recovering the class-specific means for the distal outcome. The multinomial model with K classes has $2K - 2$ model parameters and those are estimated to fit as well as possible to the conditional probabilities $P(C|X)$. Ultimately however, the best multinomial model is estimated to fit the data well and because the conditional mean $E(X|C)$ is essentially a first-order sample statistic we can generally expect that this statistic will be fitted well by the model. This is exactly what the earlier simulations illustrate. Even when the multinomial model is not correct the basic sample statistics could be fitted well. Other simulations, not reported here, also confirmed that. For example, generating data from a model where $\log(X)$ is the true predictor in the multinomial regression rather than X, which is clearly a multinomial model misspecification, did not yield bias in the the conditional mean estimates $E(X|C)$. This, however, might not always happen.

Consider the following example. We generate data as in the previous section with the exception that the mean of the auxiliary variable is 0 in Class 1 and 1 in Class 2; that is, the mean of the auxiliary variable changes over classes, which was not the case in the simulation we used in the previous section. Here there is a positive association between the auxiliary variable and the latent class variable. We generate a sample of size $N = 1{,}000$. To make sure that any difference between the methods is not due to the LCA measurement model, we adopt a perfect measurement model; that is, a model where the latent class variable is measured exactly by one binary indicator. Thus the latent class variable is perfectly measured; that is, it is observed. In that case we would expect the auxiliary model estimates to be the same as the class-specific sample means for the auxiliary variable, which can be explicitly computed now because the latent class variable is observed.

In this situation, when the latent class variable is observed, both the three-step method and the PC method will always yield the correct results, regardless of how the data set is generated; that is, the estimated class-specific means will be the same as the class-specific sample means for the auxiliary variable. This, however, is not true for Lanza's method, which will need to estimate a potentially incorrect multinomial model. When estimating the auxiliary

model with our generated data set, all three methods, namely, the three-step method, the PC method and Lanza's method, yield correct results. This again illustrates that Lanza's method is fairly robust and can perform well even when the multinomial model does not hold.

Next we introduce an outlier in the data set. We add one observation in Class 1 with an auxiliary value of 100. When estimating the auxiliary model with the three methods, the three-step method and the PC method yield the correct result, whereas Lanza's method does not. In Class 1 the sample mean for the auxiliary variable, the three-step estimate, and the PC estimate are all 1.240. Lanza's method estimate is 1.433. The problem with Lanza's method in this example is the fact that the multinomial model does not fit the conditional distribution $P(C|X)$ well. This multinomial logit assumption yields biased results for the distal outcome class-specific means.

Summary

In this section we summarize the potential failures of the various distal outcome estimation methods. First, if the measurement model has low entropy, that means that the latent class variable is poorly measured and in that case all methods can be expected to fail. The second possible failure is the case when the entropy is relatively high but the latent class variable changes when the auxiliary variable is included in the analysis. In this situation the one-step and the three-step methods can fail, whereas Lanza's method and the PC method are more robust. The third possible failure is specific to Lanza's method. If the multinomial model is substantially violated, Lanza's method estimates could be biased. Fortunately this appears to be relatively rare. The one-step and three-step methods also have distributional assumptions for the distal outcome. Within each class the distribution of the auxiliary variable is assumed to be normal. If this assumption is violated the latent class might change. This was the case in the example described previously. To the list of possible failures we also need to add the fact that the PC method will tend to be biased unless the entropy is high.

CONCLUSIONS

The new three-step approach uniformly outperforms the PC approach for analyzing the relationship between a latent class variable and an auxiliary variable. If the class separation is good, the three-step approach has the same efficiency as the one-step approach. Our simulations seem to indicate that an entropy level of 0.6 or higher provides sufficiently good class separation and in that case we can expect the three-step approach to work as efficiently as the one-step approach. In principle, the one-step approach can be used in practical applications as well. However, if the latent classification changes dramatically when the auxiliary variables are included in the model a detailed analysis should be conducted to determine the cause of the classification shift. We illustrated also that if the auxiliary variables are dependent variables, as in the case of the distal outcome, both the one-step and the three-step approach can fail due to the change in the formation of the latent classes. In the case of the distal outcome auxiliary model, Lanza's method provides a good alternative to the one-step and the three-step approaches because it will preserve the latent class variable.

In the M*plus* implementation of the three-step methods, multiple predictor variables can be used for the latent class variable and the estimated multinomial model in the third step will include all of the predictor variables. Multiple distal auxiliary variables can also be used, but the distal outcome models are estimated one at a time. The M*plus* automatic implementation for the auxiliary variables is limited to the distal outcome model and the latent class predictor model. Other models might be of interest as well, such as, for example, a distal outcome model where the distal outcome is regressed on the latent class variable and other observed variables. For such models, it is easy to manually set up all the steps of the three-step estimation method following the description provided here. The three-step procedure can be used with an arbitrary auxiliary model. The examples we presented in this article used an LCA model as a measurement model for the latent class variable. The M*plus* implementation, however, is very flexible and can use any other latent class model as the measurement model, including GMMs and any type of dependent variables.

Lanza's method as implemented in M*plus* can accommodate continuous and categorical distal outcomes, but, it is more limited in terms of the scope of models it can accommodate. The latent class measurement model can be an arbitrary measurement model but the model cannot include latent class predictors. Also, Lanza's method cannot be used with an arbitrary auxiliary model. It is important to note, however, that the underlying assumption of Lanza's method, namely, that the auxiliary model can be estimated indirectly by assuming a multinomial regression model between the latent class variable and the auxiliary variable, does not appear to have substantial drawbacks. That is, even when the multinomial regression model does not hold, in most situations Lanza's method still yields unbiased estimates.

ACKNOWLEDGMENTS

We thank Zsuzsa Bakk and Margot Bennink for uncovering an error in an earlier version of this article.

REFERENCES

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association betwen latent class membership and external variables using bias adjusted three-step approaches. In T. F. Liao (Ed.), *Sociological methodology* (pp. 272–311). Thousand Oaks, CA: Sage.

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

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

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

M*plus* Technical Appendices: Wald Test of Mean Equality for Potential Latent Class Predictors in Mixture Modeling. (2010). Retrieved from http://www.statmodel.com/download/meantest2.pdf

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Newbury Park, CA: Sage.

Muthén, B., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs. (Eds.), *Longitudinal data analysis*, (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463–469.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

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

Wang, C.-P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, *100*, 1054–1076.