Auxiliary Variables in Mixture Modeling: A 3-Step Approach Using Mplus

Tihomir Asparouhov and Bengt Muthén

Mplus Web Notes: No. 15

Version 6, February 7, 2013

This version corrects errors in the October 4, 2012 version.
These errors are corrected in the forthcoming
Mplus Version 7.1.

1 Introduction

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 analysis will be discussed here. The first type of analysis is using the latent categorical variable as a predictor of another observed variable which we call a distal outcome. The second type of analysis is when we use the observed variable as a predictor of the latent categorical variable which we call the latent class regression analysis. The standard way to conduct such an analysis is to combine the latent class model and the secondary model, such as the distal outcome model or the latent class regression model into one joint model which can be estimated with the maximum-likelihood estimator. Such an approach, however, can be flawed because the secondary model may affect the latent class formation and the latent class may lose its meaning as the latent variable measured by the indicator variables. For example, if a distal outcome variable is modeled as a normally distributed variable but it has a bimodal distribution the latent class formation may end up dominated by that distal variable so that the distribution is fitted properly as a bimodal distribution and thus the latent class variable will not be formed by the original indicator variables and will not have the desired meaning. Similarly, in latent class regression analysis if the observed variable that is intended to be a predictor for the latent class has a direct effect on one of the indicator variables, including that variable as a predictor in the latent class analysis model (and ignoring the direct effect) can result in a substantial change in the way the latent class is formed and thus again the latent class variable will loose its intended meaning. Vermunt (2010) points out also other disadvantages of the 1-step, joint model estimation approach:

> 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.

To avoid all these drawbacks several methods have been developed that can independently evaluate the relationship between the latent class variable and the distal or predictor auxiliary variables. One method is to use the pseudo class method see Wang et al. (2005), Clark and Muthén (2009), and Mplus Technical Appendices: Wald Test of Mean Equality for Potential Latent Class Predictors in Mixture Modeling (2010). With this method the latent class model is estimated first, then the latent class variable is multiply imputed from the posterior distribution obtained by the 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 of the latent class is large, i.e., the class separation is large.

An alternative approach has recently been developed in Vermunt (2010) expanding ideas presented in Bolck et al. (2004). In this approach the latent class model is estimated first. In the second step the most likely class variable S is created using the latent class posterior distribution obtained during the LCA estimation, i.e., for each observation, S is set to be the class c for which $P(C = c|U)$ is the largest, where U represents the latent class indicators. In Mplus this variable is automatically created using the SAVEDATA command with the option SAVE=CPROB. We then compute the classification uncertainty rate for S as follows

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

where N_{c_1} is the number of observations classified in class c_1 by the most-likely class variable S, S_i is the most likely class variable for the i-th observation, C_i is the true latent class variable for the i-th observation and U_i

Figure 1: Classification uncertainty rate for most likely class variable.

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

represents the class indicator variables for the i-th observation. The probability $P(C_i = c_2|U_i)$ is computed from the estimated LCA model. In Mplus 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 Most Likely Latent Class Membership (Row) by Latent Class (Column)". For example in the case of a 3 class model the probability p_{c_1,c_2} would look like in Figure 1, where the p_{c_1,c_2} is in row c_1 and column c_2. We can then compute the probability

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

where N_c is the number of observations classified in class c by the most-likely class variable S. This shows that S can be treated as an imperfect measurement of C with measurement error defined by q_{c_1,c_2}.

In the third step the most likely class variable is used as latent class indicator variable with uncertainty rates prefixed at the probabilities q_{c_1,c_2} obtained in step two. This way the measurement error in the most likely class S is taken into account in the third step model estimation. In this final stage we also include the auxiliary variable. More details on this approach are available in Vermunt (2010) where it is referred as Modal ML. Here we will refer to this method as the 3-step approach. In the Vermunt (2010) article this 3-step approach was used for latent class predictors. In this article we extend the method also for distal outcomes.

Finally in our comparisons we will also use the estimation of the joint model which includes the latent class model as well as the auxiliary variable

model. This model would in principle be expected to be the most efficient within a properly specified simulation study. However as we noted above it may in practical applications be difficult to utilize because including the auxiliary variable in the model changes the latent class model. We will call this approach the 1-step approach.

All of the above methods can easily be obtained in the Mplus 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 3-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 3-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 3-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 may 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 may lead to biases in the estimates and the standard error if the equal variance assumption is violated.

In Section 2 we present simulation studies with a distal outcome auxiliary variable and in Section 3 we present simulation studies with a predictor auxiliary variable. Section 4 presents simulation studies to evaluate the performance of the 3-step procedure in the presence of direct effect in the latent class measurement model. In Section 5 we describe a general method for estimating an arbitrary auxiliary model with a latent class variable. In Section 6 we discuss 3-step estimation for the latent transition analysis model. Section 7 concludes. In the Appendices we provide the Mplus inputs used for the above analyses.

2 Simulation study with a distal auxiliary variable

In this simulation study we estimate a 2-class model with 5 binary indicator variables. The distribution for each binary indicator variable U is determined by the usual logit relationship

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

where C is the latent class variable which takes values 1 or 2 and the threshold value τ_c is the same for all 5 binary indicators. In addition we set $\tau_2 = -\tau_1$ for all five indicators. We choose three values for τ_1 to obtain different level of class separation/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 43% and 57%. In addition to the above latent class model we also generate a normally distributed distal auxiliary variable with mean 0 in class one and mean 0.7 in class 2 and variance 1 in both classes. We apply the PC, the 3-step and the 1-step approaches to estimate the mean of the auxiliary variable in the two classes. Table 1 presents the results for the mean of the auxiliary variable in class 2. We generate 500 samples of size 500 and 2000 and analyze the data with the three methods. It is clear from the results in Table 1 that the 3-step procedure outperforms the PC procedure substantially in terms of bias, mean squared error and confidence interval coverage. When the 3-step procedure is compared to the 1-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 3-step procedure also provides good confidence interval coverage. The effect of the sample size appears to be negligible in the sample size range 500-2000. Further simulation studies are needed to evaluate the performance of the 3-step procedure for much smaller or much larger sample sizes. Appendix A contains an input file for conducting a simulation study with a distal auxiliary variable.

Next we conduct a simulation study to compare the performance of the two different 3-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 while the second approach estimates different

Table 1: Distal outcome simulation study: Bias/Mean Squared Error/Coverage

N	Entropy	PC	3-step	1-step
500	0.7	.10/.015/.76	.00/.007/.95	.00/.006/.94
500	0.6	.16/.029/.50	.01/.008/.94	.00/.007/.94
500	0.5	.22/.056/.24	.03/.017/.86	.01/.012/.96
2000	0.7	.10/.011/.23	.00/.002/.93	.00/.002/.93
2000	0.6	.15/.025/.03	.00/.002/.93	.00/.002/.94
2000	0.5	.22/.051/.00	.00/.004/.91	.00/.003/.94

means but equal variances. The second approach is more robust and more likely to converge but may suffer from the misspecifcation that the variances are equal in the different classes. We use the same simulation as above 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 2. It is clear from these results that the unequal variance 3-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 may 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 accuracy advantage for the unequal variance 3-step approach.

Table 2: Distal outcome simulation study. Comparing equal and unequal variance 3-step methods: Bias/Mean Squared Error/Coverage

N	Entropy	3-step equal variance	3-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
2000	0.7	.05/.040/.92	.00/.027/.92
2000	0.6	.09/.056/.92	.00/.027/.93
2000	0.5	.11/.094/.95	.00/.029/.92

3 Simulation study with a latent class predictor auxiliary variable

We replicate the simulation study from the previous section with the exception that the auxiliary variable is now generated as a standard normal variable and is a predictor of the latent class variable through the multinomial logistic regression

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

where $\alpha = 0.3$ and $\beta = 0.5$. We use again the three different levels for the threshold and the two different sample sizes. We generate again 500 samples and analyze the data using the three different methods. Table 3 contains the results of the simulation study for the regression coefficient β. The 3-step procedure again outperforms the PC procedure substantially in terms of bias, mean squared error and confidence interval coverage. The loss of efficiency of the 3-step procedure when compared to the 1-step method is minimal. The 3-step procedure also provides good coverage in all cases. The effect of sample size appears to be negligible here as well within the sample size range used in the simulation study. Further simulation studies are needed to evaluate the performance for much smaller or much larger sample sizes. Appendix B contains an input file for conducting a simulation study with a latent class predictor auxiliary variable.

Table 3: Latent class predictor simulation study: Bias/Mean Squared Error/Coverage

N	Entropy	PC	3-step	1-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
2000	0.7	.13/.019/.24	.00/.004/.93	.00/.004/.94
2000	0.6	.20/.042/.01	.00/.004/.95	.00/.004/.94
2000	0.5	.29/.085/.00	.01/.007/.94	.01/.006/.95

4 Simulation study with omitted direct effects from the latent class predictor auxiliary variable

In this section we study the ability of the 3-step approach to absorb misspecifications in the measurement model due to omitted direct effects from a covariate. Vermunt (2010) suggests that the 3-step estimation might be a more robust estimation method in that context. We consider 3 different situations: direct effects in LCA, direct effects in Growth Mixture Models (GMM) and direct effects in the distal outcome model.

4.1 Direct effects in LCA

The setup for this simulation study is the same as in the previous section however we generate data 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 above 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 non-zero direct effects. All non-zero direct effects γ_{pc} are set to 1. We generates different samples with L direct effects for $L = 1, 2, ..., 5$. All

non-zero direct effects are in class one. 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 2000. The generated data are analyzed with 3 different methods. Method 1 ignores the direct effect in the LCA measurement model and analyzes the regression of C on X using the 3-step procedure. Method 2 includes the direct effect in the LCA measurement model and analyzes the regression of C on X using the 3-step procedure. Method 3 is the 1-step approach which 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 3-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 3-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 can not 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 which 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 are a sufficient number of omitted direct effects the latent class variable can not be measured well only by the indicator variables. That in turn leads to substantial biases in the C on X regression using the 3-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 which includes the direct effects performs much better, however biases are found with this 3-step method as well when the entropy is 0.6. In contrast, the 3-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 may lead to misestimation of some of the parameters which in turn will lead to biases in the formation of the latent

Table 4: LCA with direct effects: absolute bias and coverage

Number of direct effects	Entropy	Method 1 3-step excluding direct effects	Method 2 3-step including direct effects	Method 3 1-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)

classes and biases in the auxiliary model estimation. To estimate Method 2 in Mplus the covariate X has to be used in the model as well as in the AUXILIARY option. In Mplus Version 7 this will not be allowed, although within a Montecarlo simulation it is allowed. To easily estimate Method 2 the covariate should be duplicated using the DEFINE command and the duplicate variable should be used in the model. This approach is illustrated in Appendix C.

The 1-step approach performs well in all cases. This finding indicates that the 3-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 3-step approach. In the above 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 3-step methods for various levels of direct effect.

4.2 Direct effects in growth mixture models

The impact of direct effects on the 3-step estimation can also be seen in the context of growth mixture models when the direct effect is not on the observed variables but it is on the growth factors. Consider the following growth mixture model (GMM).

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

where Y_t are the observed variables and I and S are the growth factors which 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. The above model simply postulates that the latent classes are determined by the pattern of growth trajectory, i.e., the latent class variable determines the mean of the intercept and the slope growth factors, but individual variation is allowed. The above growth mixture model is essentially the measurement model for the latent class variable C. In this situation we are again interested in estimating with the 3-step approach the relationship between C and X independently of the measurement model, i.e., we want to estimate the logistic regression model

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

We generated 100 samples of size 5000 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 $t = 0, 1, ..., 4$. 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, i.e., $\beta_{1c} = \beta_{2c} = 0$. Type 2 uses the same direct effects across the two classes $\beta_{1c} = 1$ and $\beta_{2c} = 0.2$, i.e., 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 3-step method that uses only the growth model as the measurement model, Method 2 use the growth

Table 5: GMM 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)

model as the measurement model but includes the direct effects from X to the growth factors. Method 3 is the 1-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 to 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, however, 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 ML estimator for the correctly specified model.

The identification of the latent class variable is more complicated in the GMM model than in the LCA model. The local independence assumption of the LCA model is not present in the GMM model. Nevertheless we see the same pattern, if the covariates have direct effects on the measurement model, these effects should be included for the 3-step approach to work well. More simulation studies are needed to evaluate the impact of the size of the direct effects on the 3-step estimation.

4.3 Direct effects for distal outcomes

In the case of the distal outcome auxiliary model, the distal outcome may have a direct effect from a covariate as well as an effect from the latent class variable. However, this direct effect will not affect the latent class

measurement model. Instead, this direct effect is a part of the auxiliary distal outcome model and it should be included in the auxiliary model. In Mplus this can not be done automatically, however the following section illustrates how this more elaborate auxiliary model can be estimated in Mplus with the 3-step procedure.

5 Using Mplus to conduct the 3-step procedure with an arbitrary secondary model

In many situations it would be of interest to use the 3-step procedure to estimate a more advanced secondary model that includes a latent class variable. In Mplus, the 3-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 3-step procedure has to be implemented manually, meaning that each of the 3 steps is performed separately. In this section we illustrate this manual 3-step estimation procedure with a simple auxiliary model where the latent class variable is a moderator for a linear regression model. 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 3-class latent 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

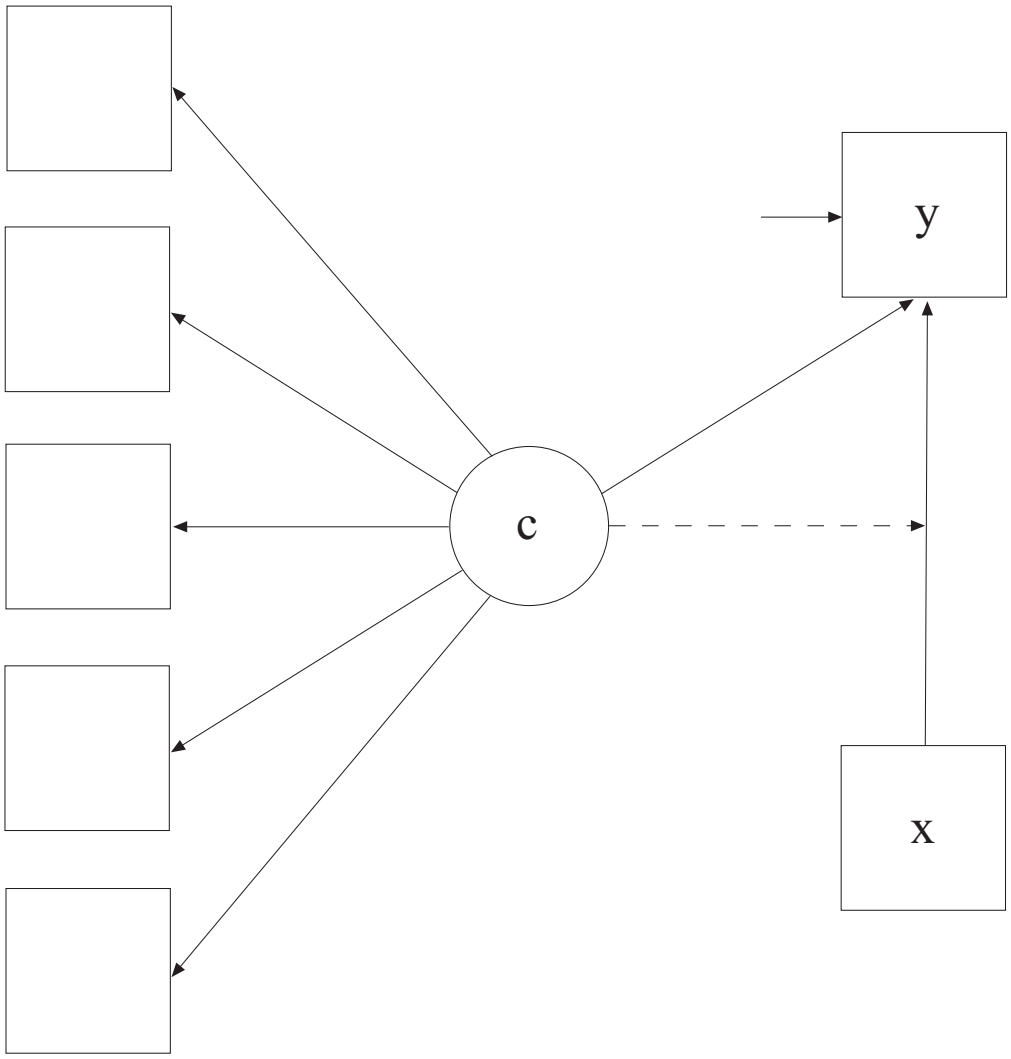
$$Y = \alpha_c + \beta_c X + \varepsilon$$

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

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

for $p = 1, ..., 10$ and $c = 1, ..., 3$. We generate a sample of size 1000 using equal classes and the following parameter values $\tau_{1p} = -1$, $\tau_{2p} = 1$, $\tau_{3p} = 1$ for $p = 1, ..., 5$, $\tau_{3p} = -1$ for $p = 6, ..., 10$. The parameters in the secondary

Figure 2: Linear regression auxiliary model



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 $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = -1$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, $\beta_2 = 0$. Appendix D contains the input file for generating this data set. Note that in this input file we don't need a model statement because we only use this input file to generate data.

The first step in the 3-step estimation procedure is to estimate the measurement part of the joint model, i.e., the latent class model. Thus in step 1 we estimate the LCA model with the 10 binary indicator variables and without the secondary model. The input file for this estimation is given in Appendix E. Note here that the Model statement is not needed. We have included that however so that the order of the classes remains the same as in the data generation. This is done just to make easy comparison between the true and the estimated parameters. In a practical application if the measurement part is an LCA model, the Model section of this input can be removed. Note also that we specified the number of random starting values to be 0 in the ANALYSIS command with the option STARTS. This is again done to avoid class order switching between the data generation procedure and the estimation procedure. This option should not be used in a practical application setting. Finally we need to clarify the use of the AUXILIARY option in the VARIABLE command. This use of the AUXILIARY option is completely different from the ones discussed in the previous sections. In this situation we do not specify a type for the auxiliary variables such as (R3STEP) or (DU3STEP). This means that the auxiliary variables are not used in the estimation. They are only included in the SAVEDATA file which will be used in the following steps. The SAVEDATA command is also used in this input file with the option SAVE=CPROB. This option produces 2 types of outputs. It produces the posterior class probabilities for each observation, which we don't actually need, as well as the most likely class variable N that we will use as a latent class indicator in the final stage estimation.

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 3x3 table titled: **Average Latent Class Probabilities for Most Likely Latent Class Membership (Row) by Latent Class (Column)**, see Figure 3. If the variable N was a perfect indicator for the true variable C and the measurement LCA model measured the latent class variable without any error, then this 3x3 matrix would have been the identity matrix with all entries on the diagonal 1 and all off-diagonal entries of 0. Unless the classes are per-

fectly and clearly separated, N will be an imperfect indicator/measurement of C. This measurement error has to be properly accounted for. Otherwise, if N is used as if it was the actual latent class variable C, the parameter estimated in the secondary model will be biased. In this step 2 of the estimation we use a calculator or Excel to compute the probabilities q_{c_1,c_2} using formula (1). The number of observations classified in the three classes can be found in the section of the Mplus output labeled **CLASSIFICATION OF INDIVIDUALS BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP**. In this example $N_1 = 346$, $N_2 = 306$, $N_3 = 348$. Also using p_{c_1,c_2} from Figure 3 and formula (1) we compute the probabilities $q_{11} = 0.829$, $q_{12} = 0.072$, $q_{13} = 0.099$, $q_{21} = 0.046$, $q_{22} = 0.811$, $q_{23} = 0.094$, $q_{31} = 0.124$, $q_{32} = 0.117$, $q_{33} = 0.807$.

Now we compute the log ratios for each class c between each category and the last category $\log(q_{i,c}/q_{3,c})$. In our example we compute

$$log(0.829/0.124) = 1.900$$

$$log(0.046/0.124) = -0.992$$

$$log(0.072/0.117) = -0.486$$

$$log(0.811/0.117) = 1.936$$

$$log(0.099/0.807) = -2.098$$

$$log(0.094/0.807) = -2.150$$

The natural logarithmic function is used in the above computation rather than the base 10 logarithmic function which in many software products and calculators is referred to as LN(). If the last category probability is estimated as 0 one can replace that probability with 0.0001 to avoid the problem of dividing by 0.

The final third step in the 3-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 computed in Step 2. The input file for our example is provided in Appendix F. Note that in this step we use the data file obtained from the SAVEDATA command in Step 1. The most likely class variable is

Figure 3: Linear regression auxiliary model

```
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
```

Table 6: Final estimates from the manual 3-step estimation with linear regression auxiliary model.

Parameter	True Value	Estimate	Standard Error
α_1	0	0.025	0.068
β_1	0.5	0.493	0.068
α_2	1	1.076	0.068
β_2	-0.5	-0.452	0.062
α_3	-1	-1.078	0.068
β_3	0	0.094	0.058

specified as a nominal variable and all the parameters $[N\#i]$ of the conditional distribution $[N|C]$ are fixed to the log ratios computed in Step 2. The parameters $[N\#1]$ and $[N\#2]$ in class 1 are fixed to the log ratios obtained from row 1 in the measurement error table: 1.900 and -0.992. The parameters $[N\#1]$ and $[N\#2]$ in class 2 are fixed to the log ratios obtained from row 2 in the measurement error table etc. In this third step we also specify the auxiliary model. In our example this is just a simple linear regression model. The estimates obtained in this final stage are presented in Table 6. These estimates are very close to the true parameter values and we conclude that the 3-step procedure works well for this example. This example also illustrates how Mplus can be used to estimate an arbitrary auxiliary model with a latent class variable in a 3-step procedure where the measurement model for the latent class variable is estimated independently of the auxiliary model.

6 Estimating latent transition analysis using the 3-step approach

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 3-step estimation can be conducted for the LTA model with Mplus 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 very desirable in the LTA context because the 1-step approach has the drawback where an observed measurement at one point in time affects the definition of the latent class variable at another point in time. The estimation is conducted manually step by step as described in the previous section. We illustrate the estimation with two different examples. The first example is a simple LTA model with 2 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 below 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.

6.1 Simple LTA

For illustration purposes we consider an example with 2-latent class variables C_1 and C_2 each measured by 5 binary indicators. The coefficient of interest, estimated in the 3-step approach is the regression coefficient of C_2 on C_1. We include four input files in Appendices G, H, I, J to illustrate the entire process.

The input file in Appendix G is used to generate data according to the true LTA model. The input file in Appendix H is used to estimate the LCA measurement model for the first class variable C_1 and to obtain the most likely class variable N_1 which will be used in step 3 as a C_1 indicator. The measurement error for N_1 is computed using the log ratios as in Section 5. The input file in Appendix I is used to estimate the LCA measurement model for the second class variable C_2 and to obtain the most likely class variable N_2 which will be used in step 3 as a C_2 indicator. The measurement error for N_2 is computed using the log ratios as in Section 5. In practical

applications both Appendices H and I do not need a model statement. We provide model statements here simply to order the classes according to the way we generated the data. The final third step is to estimate an LTA model where the variable N_1 is used as a class indicator variable for the first latent variable with prefixed error rates and the variable N_2 is used as a class indicator variable for the second latent class variable with prefixed error rates. This input file is included in Appendix J.

The 3-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, i.e., the estimate is close to the true value. Simulations studies are currently not very easy to conduct in Mplus 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 3-step estimator performs well for the LTA model. The above approach can also be used for 3-step LTA estimation with more than 2 latent class variables and also with covariates which will be used only in the third step.

6.2 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 invariant across time. The approach illustrated in Appendices G-J is inadequate and can not be used to estimate the 3-step LCA with measurement invariance because the LCA at the different time points are estimated in different input files. It is possible however to estimate 3-step LTA with measurement invariance and we illustrate that with Appendices K-O. We also illustrate in these Appendices how to include a covariate in the 3-step LTA estimation.

Appendix K contains the input file needed to generate the LTA data with a covariate. Appendix L contains the input file where the two LCA models at the two time points are estimated in parallel 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 3-step methodology. The

actual regression of C_2 on C_1 will be estimated in the last step of the 3-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 sperate files. This is because without the measurement invariance the log-likelihood of the joint model will split in two independent parts that can be estimated separately.

Note that in Appendix L we request the OUTPUT option SVALUES which provides the model input commands for the next two input files. The SVALUES output contains the final results of the model estimation formatted as an input file. At this point in the SVALUES output one has to replace the * symbol with the @ symbol because in the next two inputs we are holding the parameters fixed to the results of the joint LCA estimation from Appendix L. Appendix M contains the LCA estimation for the C_1 variable separately. With this input we obtain the most likely class variable N_1 and its measurement error. Appendix N contains the LCA estimation for the C_2 variable separately. With this input we obtain the most likely class variable N_2 and its measurement error. Note again that all the parameters in Appendices M and N are held equal to those parameters obtained in Appendix L. At this point, in step 2, we manually calculate the log ratios from the error tables for N_1 and N_2 as we did in Section 5. Appendix O contains the final third step in this estimation where N_1 and N_2 are used as C_1 and C_2 indicators with parameters fixed at the step 2 log ratios. This input now contains the auxiliary model which contains the regression of C_2 on C_1 as well as the regression of C_1 and C_2 on X.

In this particular example the true value for C_1 on C_2 is 0.5 and the 3-step estimate for that parameter is 0.63(0.19). The true value for C_2 on X is -0.5 and the 3-step estimate is -0.58(0.07). The true value for C_2 on X is 0.3 and the 3-step estimate is 0.22(0.08). All parameters of the auxiliary model are covered by the confidence intervals obtained by the 3-step estimation procedure and thus we conclude that the 3-step procedure works well for the LTA model with measurement invariance.

7 Conclusion

The new 3-step approach outperforms uniformly the pseudo-class approach for analyzing the relationship between a latent class variable and an auxiliary variable independently of the latent class model estimation. If the class separation is good the 3-step approach has the same efficiency as the 1-step approach. Our simulations seem to indicate that entropy level of 0.6 or higher is sufficiently good class separation and in that case we can expect the 3-step approach to work as efficiently as the 1-step approach. In principle the 1-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. Detailed analysis can be conducted using model modification indices for example as well as other model diagnostic tools.

In the Mplus implementation of the 3-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, however the distal outcome models are estimated one at a time. The Mplus automatic implementation for the auxiliary variables is limited to the distal outcome model and the latent class predictor model. Other models may 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 3-step estimation method following the description provided here. The 3-step procedure can be used with an arbitrary auxiliary model. The examples we presented in this paper used an LCA model as a measurement model for the latent class variable. The Mplus implementation however is very flexible and can use any other latent class model as the measurement model including for example growth mixture models and any type of dependent variables.

8 Acknowledgement

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

9 Appendix A: Input file for conducting a simulation study with a distal outcome

```
Montecarlo:
Names are u1-u5 y;
Generate = u1-u5(1);
Categorical = u1-u5;
Genclasses = c(2);
Classes = c1(2);
Nobservations = 500;
Nreplications = 500;
Auxiliary = y(DU3STEP);

Analysis: Type = Mixture;

Model Population:
%Overall%
[y@0];
y@1;
[c#1*0.3];
%c#1%
[u1$1-u5$1*-1.25];
[y*0];
%c#2%
[u1$1-u5$1*1.25];
[y*0.7];

Model:
%Overall%
[c1#1*0.3];
[y] (1); y (2); ! This command is needed so that the LCA model
                ! is estimated with no influence from the distal
                ! variable on the class formation
%c1#1%
[u1$1-u5$1*-1.25];
%c1#2%
[u1$1-u5$1*1.25];
```

10 Appendix B: Input file for conducting a simulation study with a latent class auxiliary predictor

```
Montecarlo:
Names are u1-u5 x;
Generate = u1-u5(1);
Categorical = u1-u5;
Genclasses = c(2);
Classes = c1(2);
Nobservations = 500;
Nreplications = 500;
Auxiliary = x(R3STEP);

Analysis: Type = Mixture;

Model Population:
%Overall%
[x@0];
x@1;
[c#1*0.3];
c#1 on x*0.5;
%c#1%
[u1$1-u5$1*-1.25];
%c#2%
[u1$1-u5$1*1.25];

Model:
%Overall%
[c1#1*0.3];
c1#1 on x@0; ! This command is needed so that the LCA model
             ! is estimated with no influence from the predictor
             ! variable on the class formation
%c1#1%
[u1$1-u5$1*-1.25];
%c1#2%
[u1$1-u5$1*1.25];
```

11 Appendix C: Input file for a 3-step analysis with an auxiliary variable used also as a covariate

```
variable:
Names are u1-u10 x;
usevar are u1-u10 x x2;
Categorical = u1-u10;
Classes = c(2);
Auxiliary = x(R3STEP);

define: x2=x; ! duplication of variable

data: file=dup3st.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c#1*0.3];
u1 on x2*0;

%c#1%
[u1$1-u10$1*-0.75];
u1 on x2*1;

%c#2%
[u1$1-u10$1*0.75];
u1 on x2*0;
```

12 Appendix D: Input file for generating data for manual 3-step estimation

```
Montecarlo:
Names are u1-u10 y x;
Generate = u1-u10(1);
Categorical = u1-u10;
Genclasses = c(3);
Classes = c(3);
Nobservations = 1000;
Nrep = 1;
save=man3step.dat;

Analysis: Type = Mixture;

Model Population:

%Overall%
[x@0]; x@1;
y*1;
y on x*0;

%c#1%
[u1$1-u10$1*-1];
[y*0];
y on x*0.5;

%c#2%
[u1$1-u10$1*1];
[y*1];
y on x*-0.5;

%c#3%
[u1$1-u5$1*1];
[u6$1-u10$1*-1];
[y*-1];
y on x*0;
```

13 Appendix E: Input file for step 1 in the 3-step estimation

```
variable:
Names are u1-u10 y x;
Categorical = u1-u10;
Classes = c(3);
usevar are u1-u10;
auxiliary=y x;

data: file=man3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%

%c#1%
[u1$1-u10$1*-1];

%c#2%
[u1$1-u10$1*1];

%c#3%
[u1$1-u5$1*1];
[u6$1-u10$1*-1];

SAVEDATA: FILE= man3step2.dat; SAVE=CPROB;
```

14 Appendix F: Input file for step 3 in the 3-step estimation

```
variable:
Names are u1-u10 y x p1-p3 n;
usevar are y x n;
classes = c(3);
nominal=n;

data: file=man3step2.dat;

Analysis: Type = Mixture; starts=0;

Model:

%overall%
Y on X;

%C#1%
[N#1@1.900];
[N#2@-0.992];
Y on X;

%C#2%
[N#1@-0.486];
[N#2@1.936];
Y on X;

%C#3%
[N#1@-2.098];
[N#2@-2.150];
Y on X;
```

15 Appendix G: Input file for LTA data generation

```
Montecarlo:
Names are u11-u15 u21-u25;
Generate = u11-u15(1) u21-u25(1);
Categorical = u11-u15 u21-u25;
Genclasses = c1(2) c2(2);
Classes = c1(2) c2(2);
Nobservations = 2000;
Nrep = 1;
save=conc3step.dat;

Analysis: Type = Mixture;

Model Population:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5;

MODEL population-c1:

%c1#1%
[u11$1-u15$1*-1];

%c1#2%
[u11$1-u15$1*1];

MODEL population-c2:

%c2#1%
[u21$1-u25$1*-1];

%c2#2%
[u21$1-u25$1*1];
```

16 Appendix H: Input file for 3-step LTA analysis, estimating LCA for C_1

```
variable:
Names are u11-u15 u21-u25;
usevar are u11-u15;
Categorical = all;
Classes = c1(2);
auxiliary=u21-u25;

data: file=conc3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c1#1*0.3];

%c1#1%
[u11$1-u15$1*-1];

%c1#2%
[u11$1-u15$1*1];

savedata: file=c1.dat; save=cprob;
```

17 Appendix I: Input file for 3-step LTA analysis, estimating LCA for C_2

```
variable:
Names are u11-u15 u21-u25 p1 p2 n1;
usevar are u21-u25;
Categorical = all;
Classes = c2(2);
auxiliary=u11-u15 n1;

data: file=c1.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c2#1*0.3];

%c2#1%
[u21$1-u25$1*-1];

%c2#2%
[u21$1-u25$1*1];

savedata: file=c2.dat; save=cprob;
```

18 Appendix J: Input file for 3-step LTA analysis, estimating the final auxiliary model

```
variable:
Names are u21-u25 u11-u15 n1 p1 p2 n2;
usevar are n1 n2;
nominal n1 n2;
Classes = c1(2) c2(2);

data: file=c2.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5;

MODEL c1:

%c1#1%
[n1#1@1.864];

%c1#2%
[n1#1@-2.138];

MODEL c2:

%c2#1%
[n2#1@1.841];

%c2#2%
[n2#1@-1.842];
```

19 Appendix K: Input file for LTA data generation with measurement invariance and a covariate

```
Montecarlo:
Names are u11-u15 u21-u25 x;
Generate = u11-u15(1) u21-u25(1);
Categorical = u11-u15 u21-u25;
Genclasses = c1(2) c2(2);
Classes = c1(2) c2(2);
Nobservations = 2000;
Nrep = 1;
save=conc3step.dat;

Analysis: Type = Mixture;

Model Population:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5 x*0.3;
c1#1 on x*-0.5;
x*1;

MODEL population-c1:

%c1#1%
[u11$1-u15$1*-1];

%c1#2%
[u11$1-u15$1*1];

MODEL population-c2:

%c2#1%
```

[u21$1-u25$1*-1];

%c2#2%
[u21$1-u25$1*1];

Model:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5 x*0.3;
c1#1 on x*-0.5;

MODEL c1:

%c1#1%
[u11$1-u15$1*-1];

%c1#2%
[u11$1-u15$1*1];

MODEL c2:

%c2#1%
[u21$1-u25$1*-1];

%c2#2%
[u21$1-u25$1*1];

20 Appendix L: Input file for 3-step LTA estimation with measurement invariance: step 1

```
variable:
Names are u11-u15 u21-u25 x;
Categorical = u11-u15 u21-u25;
Classes = c1(2) c2(2);
auxiliary=x;

data: file=conc3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c1#1*0.3];
[c2#1*0.3];

MODEL c1:

%c1#1%
[u11$1-u15$1*-1] (t1-t5);

%c1#2%
[u11$1-u15$1*1] (tt1-tt5);

MODEL c2:

%c2#1%
[u21$1-u25$1*-1] (t1-t5);

%c2#2%
[u21$1-u25$1*1] (tt1-tt5);

output: svalues;
```

21 Appendix M: Input file for 3-step LTA estimation with measurement invariance: step 1 for C1

```
variable:
Names are u11-u15 u21-u25 x;
usevar are u11-u15;
Categorical = all;
Classes = c1(2);
auxiliary=u21-u25 x;

data: file=conc3step.dat;

Analysis: Type = Mixture; starts=0;

Model:

%OVERALL%
[ c1#1@0.19434 ];

%C1#1%
[ u11$1@-0.97524 ] (t1);
[ u12$1@-0.98527 ] (t2);
[ u13$1@-0.96129 ] (t3);
[ u14$1@-0.97072 ] (t4);
[ u15$1@-0.89841 ] (t5);

%C1#2%
[ u11$1@1.02624 ] (tt1);
[ u12$1@1.00941 ] (tt2);
[ u13$1@1.03036 ] (tt3);
[ u14$1@1.05849 ] (tt4);
[ u15$1@1.08370 ] (tt5);

savedata: file=c1.dat; save=cprob;
```

22 Appendix N: Input file for 3-step LTA estimation with measurement invariance: step 1 for C2

```
variable:
Names are u11-u15 u21-u25 x p1 p2 n1;
usevar are u21-u25;
Categorical = all;
Classes = c2(2);
auxiliary=u11-u15 x n1;

data: file=c1.dat;

Analysis: Type = Mixture; starts=0;

Model:

%OVERALL%
[ c2#1@0.66961 ];

%C2#1%
[ u21$1@-0.97524 ] (t1);
[ u22$1@-0.98527 ] (t2);
[ u23$1@-0.96129 ] (t3);
[ u24$1@-0.97072 ] (t4);
[ u25$1@-0.89841 ] (t5);

%C2#2%
[ u21$1@1.02624 ] (tt1);
[ u22$1@1.00941 ] (tt2);
[ u23$1@1.03036 ] (tt3);
[ u24$1@1.05849 ] (tt4);
[ u25$1@1.08370 ] (tt5);

savedata: file=c2.dat; save=cprob;
```

23 Appendix O: Input file for 3-step LTA estimation with measurement invariance: step 3

```
variable:
Names are u21-u25 u11-u15 x n1 p1 p2 n2;
usevar are n1 n2 x;
nominal n1 n2;
Classes = c1(2) c2(2);

data: file=c2.dat;

Analysis: Type = Mixture; starts=0;

Model:

%Overall%
[c1#1*0.3];
[c2#1*0.3];
c2#1 on c1#1*0.5 x*0.3;
c1#1 on x*-0.5;

MODEL c1:

%c1#1%
[n1#1@1.925];

%c1#2%
[n1#1@-2.020];

MODEL c2:

%c2#1%
[n2#1@1.787];

%c2#2%
[n2#1@-2.084];
```

References

[1] 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.

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

[3] Mplus Technical Appendices: Wald Test of Mean Equality for Potential Latent Class Predictors in Mixture Modeling (2010) http://www.statmodel.com/download/meantest2.pdf

[4] 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 Publications.

[5] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York, J. Wiley & Sons.

[6] Vermunt, J. K. (2010) Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. Political Analysis, 18, 450-469.

[7] Wang C-P, Brown CH, 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.