

# Expanding the Bayesian Structural Equation, Multilevel and Mixture Models to Logit, Negative-Binomial and Nominal Variables

*Tihomir Asparouhov & Bengt Muthén*

Version 3

November 18, 2020

## **1 Introduction**

The Bayesian estimation of the structural equation modeling framework implemented in Mplus is described in Asparouhov and Muthén (2010a). That framework includes mixture models as well as multilevel models. The observed endogenous variables can be normally distributed continuous outcomes or ordered polytomous (including binary) variables based on the probit link function. In this note we describe how the framework is expanded to include these new types of variables: count variables based on the Poisson and the Negative-Binomial distribution, Nominal variables as well as binary variables based on the logit link function. To a very large extent, this expanded framework is the same framework that is estimated in Mplus with the ML estimator, see Muthén and Asparouhov (2007). The model is largely unchanged but we can now utilize the efficiency of the Bayesian estimation. With the ML estimator, the number of latent variables and random effects can not exceed 3 or 4 because the computational speed grows exponentially with the number of latent variables and random effects. Numerical integration is typically required in the ML estimation of latent variable models with non-normal outcomes. In the Bayesian estimation, the computational speed grows linearly with the number of latent variables, and generally, an unlimited number of latent variables can be used without substantially increasing the computational time.

Categorical variables in the Asparouhov and Muthén (2010a) framework, based on the probit link function, utilize the conceptualization of an underlying latent variable. For each categorical variable  $Y$ , there is a normally distributed latent variable  $Y^*$  that is cut according to certain threshold parameters. If  $Y^*$  is smaller than the first threshold, then  $Y$  takes the first category. If  $Y^*$  is between the first and the second threshold, then  $Y$  takes the second category and so on. The existence of such underlying latent variable is immediately provided by the probit link function. The latent variable construct is very important in the Bayesian estimation. Any structural model that can be formulated for  $Y$ , can be formulated as a structural model for  $Y^*$ . Because  $Y^*$  is normally distributed, all conditional distributions (for structural parameters, latent factors, random effects and missing data) in the MCMC estimation are explicit. Provided with conjugate priors, the MCMC estimation is fast and efficient.

For other types of variables such as Count, Logit-categorical or Nominal, a conceptualization of such underlying normal variables had not been found until very recently. Fox's (2010) work on item response modeling, for example, in the absence of underlying latent variables, utilizes the Metropolis-Hastings algorithm as a part of the MCMC algorithm. Such an approach tends to be more computationally demanding, less efficient than explicit conditional distributions, more difficult to implement in a generalized framework, and more prone to slow mixing estimation. Groundbreaking recent work by Pillow and Scott (2012) and Polson, Scott, and Windle (2013) yields a normally distributed underlying variable methodology for Negative-Binomial count variables and Logit-Binary variables. Both methods utilize a technique based on the Polya-Gamma (PG) distribution. The approach easily extends to Poisson variables as well as nominal variables. This new method is uniquely suited for structural and multilevel models and is now implemented in the Mplus version 8.5. With the PG method, regression parameters with conjugate normal priors can be updated in the MCMC estimation from their conditional normal distributions. Similarly, normally distributed latent variables or random effects (random slopes, loadings or intercepts) can be updated from explicit conditional normal distributions.

An additional feature that is now also implemented in Mplus is the multinomial logistic regressions for latent class variables. Latent class variables in Mplus are generally modeled as nominal variables and in prior Mplus versions the latent class variable regression on other variables was not available with the Bayesian estimation. This is now resolved in Mplus 8.5 with the

PG methodology.

Missing data on Count/Nominal/Categorical variables is typically fairly easy to resolve in likelihood based methods (Bayes and ML), however, missing data on the predictors/covariates of such variables is not. In the ML estimation for example, such missing data leads to additional dimensions of numerical integration. Even for a simple regression model, the dimensions of numerical integration could easily become substantial depending on how much missing data there are in the covariates. In this situation the PG methodology helps as well and the missing values can be generated in the MCMC estimation from explicit conditional distributions. This also extends to the situation where latent class variables have missing predictors and can even be applied to 3-stage estimation, see Asparouhov and Muthén (2014), where the final step of the estimation is conducted with the Bayes estimator instead of the ML estimator.

There is one drawback of the PG methodology that makes the Bayesian estimation slower, however. If categorical data is modeled with the probit link function, the underlying normal variable has a conditional distribution that varies linearly with the model covariates. The model parameters remain the same across individuals, which makes matrix manipulations efficient. This does not apply to the PG methodology. The underlying variable for Count/Nominal/Logit variables has a conditional distribution that varies across observations. Matrix manipulation must be performed separately for each observation. This affects the computational speed of the estimation but it does not affect the generality of the methodology or the mixing efficiency.

In the next section we describe the PG methodology and how it is implemented in Mplus. We then illustrate the Bayesian estimation with several simulation studies.

## 2 Estimation

We begin by providing a formal definition for the PG distribution. A random variable  $W$  has a Polya-Gamma distribution with parameters  $b$  and  $c$ , i.e.,  $W \sim PG(b, c)$ , if  $W$  is obtained as the following infinite sum

$$W = \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{G_k}{(k - 0.5)^2 + (c/(2\pi))^2} \quad (1)$$

where  $G_k$  are independent gamma random variables with distribution  $Ga(b, 1)$ . It is not possible to simplify the above infinite sum, but in most cases the sum can be approximated as a finite sum. For example, the first 500 terms of the above sum provide a good approximation for most values of  $b$  and  $c$ . The  $PG(b, c)$  distribution takes only positive values and has mean

$$E(W) = (b/(2c)) \tanh(c/2) \quad (2)$$

and variance

$$Var(W) = (b/(4c^3))(\sinh(c) - c)/(\cosh(c/2))^2. \quad (3)$$

For large values of  $b$  and  $c$  ( $> 200$ ) the PG distribution can be approximated by a normal distribution. Also, for large values of  $c$ , the variance of the distribution goes to 0 and so the distribution becomes approximately equal to a constant. Note also that the parameter  $b$  is always positive, while the parameter  $c$  can be positive and negative and the PG distribution is independent of the sign of  $c$ .

To be able to use this distribution in the MCMC estimation, an efficient method for generating PG random variables is needed. Fortunately such a method has been established in Windle, Polson and Scott (2014). Next we describe how the PG distribution is used to facilitate the Bayesian estimation for Logistic, Negative-Binomial and Nominal regressions.

## 2.1 Logistic regression

Consider the logistic regression for a binary outcome  $Y$  (with outcome values of 0 and 1) given by the equation

$$P(Y = 0) = \frac{1}{1 + \text{Exp}(\beta X)} \quad (4)$$

where  $X$  represents a set of covariates and  $\beta$  represents a set of parameter that are to be estimated.

The underlying variable  $Y^*$  in this case is constructed in two step. In the first step we generate

$$W \sim PG(1, \beta X). \quad (5)$$

In the second step we compute  $Y^*$

$$Y^* = \frac{Y - 0.5}{\sqrt{W}} \sim N(\beta\sqrt{W}X, 1). \quad (6)$$

Since  $Y^*$  is normally distributed, the estimation of the  $\beta$  coefficients can be obtained from the implied linear regression

$$Y^* = \beta\sqrt{W}X + \varepsilon, \quad (7)$$

where  $\varepsilon$  has a standard normal distribution. Note that in this linear regression, the predictor variables are now  $\sqrt{W}X$  rather than the original covariates  $X$  used in the logistic regression. Once the underlying normal variable is generated, the logistic regression equation is replaced by the linear regression for  $Y^*$ . Thus any structural/multivariate/multilevel model involving the logistic regression for  $Y$  is transformed into a structural/multivariate/multilevel model for  $Y^*$ . Such models are then estimated as in Asparouhov and Muthén (2010a).

To be more specific, the MCMC estimation proceeds as follows. In each MCMC iteration, we generate  $W$  and compute  $Y^*$ . We then update any structural parameters, random effects, latent variables and missing data, based on the normally distributed  $Y^*$ . Note again that the covariates in the logistic equation are changed to  $\sqrt{W}X$ . If those same covariates are used as predictors in another equation, they are unchanged for that other equation or are changed according to a different PG deviate. Also note that if the covariates  $X$  must be updated due to missing values or if the covariates include latent variables, the scale coefficient  $\sqrt{W}$  is then attached to the regression parameters, i.e., in the Gibbs sampler step that updates  $X$ , the coefficients  $\beta$  in the logistic regression are replaced with the coefficients  $\beta\sqrt{W}$  in the linear regression of  $Y^*$ .

In the above description, we did not use the intercept in the logistic regression. The intercept is a special case of a regression parameter for a covariates that is the constant 1. In the Mplus implementation, however, the intercept is actually used and is called a threshold  $\tau$ . In line with the probit regression, the actual Mplus parameterization for the logistic regression is as follows

$$P(Y = 0) = \frac{\text{Exp}(\tau - \beta X)}{1 + \text{Exp}(\tau - \beta X)} = \frac{1}{1 + \text{Exp}(-\tau + \beta X)}. \quad (8)$$

Thus the logistic regression intercept is equivalent to the threshold parameter in Mplus but is with the opposite sign.

The PG methodology does not extend to non-binary ordered polytomous variables with the logit link function. The Mplus implementation allows for

the simultaneous modeling of binary variables with the Logit link function and non-binary ordered polytomous variables with the Probit link function. To use such modeling the option LINK=PROBIT LOGIT must be specified in the ANALYSIS command. The natural extension of the PG methodology to non-binary categorical variables leads to Nominal variables rather than ordered polytomous variables. We will discuss the Nominal implementation further below.

In a bivariate model where there are two binary variables  $Y_1$  and  $Y_2$  and two underlying normally distributed variables  $Y_1^*$  and  $Y_2^*$ , a natural question arises regarding the correlation between  $Y_1^*$  and  $Y_2^*$ . Such correlation can be easily estimated. However, the correlation does not naturally translate into some kind of association between the observed variables  $Y_1$  and  $Y_2$ . In the Mplus implementation, such correlations are not pursued at this time. The main tool for correlating binary variables in this logit based modeling framework is to use normally distributed latent variables that predict both binary variables (as in item response theory). This observation applies to the rest of the PG-based types of variables: Negative-Binomial, Poisson and Nominal variables.

## 2.2 Negative-Binomial regression

The Negative-Binomial regression has a number of different parameterizations. The first parameterization that we describe here is called the NB-2 parameterization, see Hilbe (2011). In the NB-2 parameterization, the probability distribution function for the Negative-Binomial variable  $Y$  is given by

$$P(Y = k) = \binom{k+r-1}{r-1} p^r (1-p)^k. \quad (9)$$

The interpretation of this distribution is that  $Y$  represents the number of successes in a sequence of Bernoulli (binary) trials before the occurrence of the  $r$ -th failure, where the parameter  $p$  represents the failure probability. Such an interpretation requires that  $r$  is an integer but in fact the Negative-Binomial distribution is defined for any positive value  $r$  using the notation that

$$\binom{k+r-1}{r-1} = \frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)} = \prod_{i=1}^k \frac{r-1+i}{i}, \quad (10)$$

where  $\Gamma$  is the Gamma function. The mean of  $Y$  is

$$\mu = E(Y) = \frac{r(1-p)}{p} \quad (11)$$

and the variance is

$$v = Var(Y) = \frac{r(1-p)}{p^2}. \quad (12)$$

Instead of using the parameter  $r$ , the Mplus implementation uses the parameter  $\alpha$

$$\alpha = \frac{1}{r} \quad (13)$$

which is called the dispersion parameter. It can be shown that

$$p = \frac{1}{1 + \alpha\mu} \quad (14)$$

$$v = \mu + \alpha\mu^2. \quad (15)$$

The above equation has a symbolic meaning as it explicates how the Negative-Binomial distribution differs from the Poisson distribution for which the mean and the variance are equal. In fact when  $\alpha$  is set to 0, the Negative-Binomial distribution becomes equivalent to the Poisson distribution. When  $\alpha > 0$ , the variance of the Negative-Binomial distribution is bigger than the mean and the term  $\alpha\mu^2$  represents the difference between the variance and the mean.

Next we discuss how the Negative-Binomial distribution is used in regression models. For the NB-2 parameterization, the relationship between  $Y$  and a set of covariates  $X$  is defined by the following equation

$$\mu = E(Y) = Exp(\beta X), \quad (16)$$

where  $\beta$  is a vector of regression parameters. The intercept is not included in this discussion again as the intercept is simply the special case of a regression parameter for the constant covariate of 1. From (16) we derive

$$p = \frac{1}{1 + \alpha\mu} = \frac{1}{1 + Exp(\beta X + \log(\alpha))}. \quad (17)$$

We denote this model by  $Y \sim NB2(\beta X, \alpha)$ .

Mplus implements also a different parameterization which we refer to as the PG parameterization. The difference between the PG parameterization

and the NB-2 parameterization is only in how the variable  $Y$  is related to the covariates  $X$ . In this alternative parameterization, equations (16-17) are replaced by

$$\mu = E(Y) = \frac{Exp(\beta X)}{\alpha} = Exp(\beta X - \log(\alpha)) \quad (18)$$

$$p = \frac{1}{1 + \alpha\mu} = \frac{1}{1 + Exp(\beta X)}. \quad (19)$$

We denote this model by  $Y \sim NBPG(\beta X, \alpha)$ .

The difference between the two parameterizations is quite simple and it is only in the intercept of the regression model. If the intercept of the NB-2 model is denoted by  $\alpha_{NB2}$  and the intercept of the NBPG model is denoted by  $\alpha_{PG}$ , the two parameters are related by the following equation

$$\alpha_{NB2} = \alpha_{PG} - \log(\alpha) \quad (20)$$

where  $\alpha$  is the dispersion parameter. The change in the parameterization has no effect on any other parameter, i.e., regression coefficients or the dispersion parameter.

In the Mplus software, the NB-2 parameterization is also implemented with the ML estimator, while the PG parameterization is only available with the Bayesian estimator. A variable can be specified as an NB-2 variable using the option `COUNT=Y(nb)`. A variable can be specified as PG variable using the option `COUNT=Y(nbpg)`. A Poisson distribution variable can be specified simply as `COUNT=Y`. The Bayesian implementation for the Poisson distribution is simply a Negative-Binomial variable, using the NB-2 parameterization, with the dispersion parameter  $\alpha$  fixed to 0.01, i.e., this implementation is an approximation to the Poisson distribution. The PG methodology does not directly extend to the Poisson distribution and it must be approximated that way as a Negative-Binomial distribution. The dispersion parameter can not be fixed directly to 0 as that will cause numerical problems in the estimation; see equation (23) below, where the log of the dispersion parameter is evaluated.

The original work by Polson, Scott and Windle (2013) uses the PG parameterization of the Negative-Binomial distribution. This is one reason why we implemented this parameterization in Mplus as well. The NB-2 parameterization was also implemented as this would be considered the most commonly used parameterization. Furthermore, simulation studies revealed that



the mixing quality of the Bayesian estimation is not the same between the two parameterizations for models that involve constraints on the intercept parameters (where the intercept parameters are held fixed to a particular value, such as in growth models, or where intercept parameters are held equal to other parameters) in the Negative-Binomial regression. Furthermore, models involving intercept constraints, such as latent growth models, are not equivalent between the two parameterizations. Intercept constraints in the NB2 parameterization translates into a different set of constraints for the NBPG parameterization. The NB-2 parameterization appears to have slightly worse mixing performance as compared to the PG parameterization but in most situations the difference would be ignorable.

### 2.2.1 The Bayesian estimation of the Negative-Binomial regression

The Bayesian estimation of the Negative-Binomial regression is very similar to that of the Logistic regression. Each Negative-Binomial variable  $Y$  is augmented with an underlying normally distributed variable  $Y^*$  and the Negative-Binomial regression model is transformed to a linear regression model for  $Y^*$ . At that point, as in the Logistic regression case, the model is estimated with the standard methodology.

First we described the construction of  $Y^*$  for the PG parameterization. The first step is to generate  $W$  as

$$W \sim PG(r + Y, \beta X). \quad (21)$$

In the second step,  $Y^*$  is computed as follows

$$Y^* = \frac{Y - r}{2\sqrt{W}} \sim N(\beta X \sqrt{W}, 1). \quad (22)$$

The construction of the underlying  $Y^*$  in the NB-2 parameterization is as follows. In the first step we generate  $W$  as

$$W \sim PG(r + Y, \beta X + \log(\alpha)). \quad (23)$$

In the second step  $Y^*$  is computed as follows

$$Y^* = \frac{Y - r}{2\sqrt{W}} - \log(\alpha)\sqrt{W} \sim N(\beta X \sqrt{W}, 1). \quad (24)$$

As in the logistic regression, in each MCMC iteration a new value for  $W$  is generated based on the current model parameters,  $Y^*$  is computed, the covariates  $X$  are multiplied by  $\sqrt{W}$  and the parameters  $\beta$  are estimated as in the linear regression estimation of  $Y^*$  on  $X\sqrt{W}$ .

### 2.2.2 Estimating the dispersion parameter

The Negative-Binomial regression has one additional parameter to be estimated: the dispersion parameter. Here we follow the approach described in Zhou and Carin (2015) and Neelon (2019) based on the Chinese restaurant table (CRT) distribution. We begin with the PG parameterization. Conditional on all model parameters, we augment the data with the following variable

$$A = \sum_{i=1}^n \sum_{j=1}^{Y_i} A_{ij} \quad (25)$$

where  $n$  is the same size,  $Y_i$  is the value of the Negative-Binomial variable for the  $i$ -th observation, and  $A_{ij}$  is a binary 0/1 variable with the following distribution

$$P(A_{ij} = 1) = \frac{r}{r + j - 1}. \quad (26)$$

If the dispersion parameter is given the conjugate Inverse-Gamma prior  $IG(a, b)$ , the conditional distribution of  $\alpha$  given all model parameters and data, including the above augmented variable  $A$ , is the following Inverse-Gamma distribution

$$[\alpha|*] \sim IG(a + A, b - \sum_{i=1}^n \log(p_i)) \quad (27)$$

where  $p_i$  is the probability given in equation (19) computed for the  $i$ -th observation.

For the NB-2 parameterization, the estimation is slightly more complex. The data augmentation for the variable  $A$  remains the same. However, the probability  $p_i$  computed in (17) for the NB-2 parameterization clearly contains the value of  $\alpha$ . Thus, the above explicit conditional distribution computation breaks down. The RHS in equation (27) should not contain  $\alpha$ . To work around this problem, we internally use  $\alpha_{PG}$  as the model parameter instead of  $\alpha_{NB2}$ . If  $\alpha_{PG}$  is the model parameter, then  $p_i$  no longer contains  $\alpha$  and is simply computed from the rest of the model parameters.

This internal reparameterization, however, implies certain limitations on the structural model that can be estimated with the CRT dispersion updating and the NB-2 parameterization. Here we list these limitations. The intercept in the Negative-Binomial regression can not be a fixed parameter since that parameter is needed for the internal reparameterization. The intercept in the Negative-Binomial regression can not be held equal to other parameters because that kind of equality constraint will not hold in the internal reparameterization. The intercept in the Negative-Binomial regression can not have an informative prior as that will also break down in the internal reparameterization.

When these model limitations apply, we use a different updating procedure for the dispersion parameter. Namely, we use the Metropolis-Hastings (MH) method, also discussed in Neelon (2019). The Mplus implementation of the MH method is as follows. At each MCMC iteration, a random deviate  $\varepsilon$  is drawn from the normal distribution  $N(0, v_0)$ . If the current dispersion parameter is  $\alpha$ , the proposal dispersion parameter  $\alpha^*$  is computed as follows

$$\alpha^* = \alpha \text{Exp}(\varepsilon). \quad (28)$$

The proposal parameter is then accepted with probability

$$\min \left( 1, \frac{P_0(\alpha^*) \prod_{i=1}^n P(Y_i|\alpha^*)}{P_0(\alpha) \prod_{i=1}^n P(Y_i|\alpha)} \right), \quad (29)$$

where  $P_0$  is the prior density function for the dispersion parameter,  $n$  is the sample size,  $Y_i$  is the  $i$ -th observed value and  $P(Y_i|\alpha)$  and  $P(Y_i|\alpha^*)$  are computed using (9) and (17).

The initial value for the proposal variance  $v_0$  is set to 0.01 and it is adjusted every 50 iterations to maintain approximate rate of acceptance of the new parameters around 50%. This variance can be adjusted only in the burnin period which by default is set to the first 1000 MCMC iterations. The proposal variance  $v_0$  does not change beyond the burnin period to preserve the validity of the MH method. Subsequently, more iterations are conducted to achieve convergence and to obtain the posterior distribution of the parameters. In addition, the first 50 iterations in this process are computed using the conjugate CRT method discussed above, which tends to give good starting values for the estimation.

The MH method can be used for any model estimation, however, the conjugate CRT method yields a more efficient mixing approach. In the CRT

method, every MCMC iteration provides a new draw from the posterior distribution of the dispersion parameter. In contrast, the MH method provides a new draw every other iteration on average. Thus, Mplus uses the CRT method, unless the model includes intercept constraints, as discussed above, in which case we use the less efficient MH method.

### 2.3 Nominal regression

Suppose that a nominal variable  $Y$  takes  $k$  possible values:  $1, \dots, k$ , which are considered unordered values. Suppose that  $X$  represents a vector of covariates. The nominal regression model for this variable is given by the following equation

$$P(Y = j) = \frac{\text{Exp}(\beta_j X)}{\sum_{i=1}^k \text{Exp}(\beta_i X)}, \quad (30)$$

where for identification purposes the  $\beta_k = 0$  and  $\beta_j$  for  $j = 1, \dots, k - 1$  are vectors of regression parameters. If  $k = 2$  the nominal regression model is the same as the logistic regression model for the binary variable discussed earlier. It turns out that the nominal regression parameters can be estimated using a sequence of  $k - 1$  logistic regressions. To estimate  $\beta_1$ , we form a new binary variable

$$Z = \begin{cases} 1, & \text{if } Y = 1 \\ 0, & \text{if } Y > 1 \end{cases} . \quad (31)$$

Given (72), we get that

$$P(Z = 0) = \frac{1}{1 + \text{Exp}(\beta_1 X + A)} \quad (32)$$

where

$$A = -\log\left(\sum_{i=2}^k \text{Exp}(\beta_i X)\right) \quad (33)$$

does not depend on  $\beta_1$ . Since equation (32) is essentially a logistic regression for  $Z$ , we can use the PG based method described earlier to compute the conditional distribution of  $[\beta_1 | \beta_2, \dots, \beta_{k-1}]$  needed for the Bayesian estimation. Similarly we can estimate  $[\beta_2 | \beta_1, \beta_3, \dots, \beta_{k-1}]$  by forming a new binary variable and logistic regression. The process is repeated  $k - 1$  times.

The above estimation method does not allow for the simultaneous estimation of  $\beta_1, \beta_2, \dots, \beta_{k-1}$ . Therefore the Bayesian estimation method can not

accommodate constraints that involve all of the nominal regression parameters. In general, this appears to be a minor limitation. Equality constraints modeling among the nominal regression parameters is rare. However, the limitation is important in the general SEM model. Two situations in particular are hindered by the sequential estimation of the model parameters. The first one is when one of the predictors has missing values. The second situation is when one of the predictors is a continuous latent variable. To update the latent variable (or equivalently the missing predictor), the regression parameters and the predictors switch roles. The regression parameters become the predictors and the predictors are estimated as regression parameters that are equal across the categories. The situation creates equality constraints across the regression parameters because the predictor is the same in all categories. This problem must be resolved with a different method. The MH method can be utilized here as well. Consider the updating a latent variable  $\eta$  which is a predictor for a set of nominal variable  $Y_1, \dots, Y_m$ . First we compute the conditional distribution of  $\eta$  given all other variables except the nominal variables. This can be computed as in Asparouhov and Muthén (2010a) and the result is a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . If the current value of the latent variable is represented by  $\eta$ , the new proposal value  $\eta^*$  is computed as

$$\eta^* = \eta + \varepsilon \quad (34)$$

where  $\varepsilon$  is drawn from a normal distribution  $N(0, v_0)$ . The proposal value is accepted with probability

$$\min \left( 1, \text{Exp}(((\eta - \mu)^2 - (\eta^* - \mu)^2)/(2\sigma^2)) \frac{\prod_{i=1}^m P(Y_i|\eta^*)}{\prod_{i=1}^m P(Y_i|\eta)} \right). \quad (35)$$

Here, the  $N(\mu, \sigma^2)$  distribution essentially takes the role of the prior. The probabilities  $P(Y_i|\eta^*)$  and  $P(Y_i|\eta)$  are computed as in (72). The same proposal distribution variance  $v_0$  is used for all observations in the sample. That variance is adjusted in every MCMC iteration during the burnin phase to maintain a rejection/acceptance ration near 50% across the entire population.

The second situation where the MH method is needed is the case when a normally distributed predictor of nominal variables has missing values. The computation in that case is similar.

Note here that the MH method is not needed when the Nominal variable is binary. The problem with equalities across the categories occurs only for

nominal variables with more than 2 categories. In the Mplus implementation it is best to specify a binary nominal variable as a logit Binary. This way the more efficient estimation with the explicit conditional distributions is utilized.

## 2.4 Missing Data

There are two distinct types of missing data that must be addressed. Missing data on the dependent Nominal/Count/Logit variables and missing data for predictors of Nominal/Count/Logit variables.

First we consider the missing data on the dependent variables. There are two different variations in the Bayesian estimation. In the first approach, the missing data is imputed within each MCMC iteration from the current model estimates. Thereafter, the imputed values are used the same way as the observed values and are essentially used to update model parameters, missing data, etc. The second approach, which is equivalent to the first approach, does not impute missing values. Instead when a quantity must be updated conditional on the dependent variable, we only condition on the actual observed values. The first approach has the advantage of somewhat simpler computations and it also provides imputed values for the missing dependent variables if such are needed. The second approach has the advantage that it mixes more efficiently as it has fewer unknown quantities to update and is computationally faster. Mplus implements the second approach. When multiple imputations are requested, however, the values are imputed as in the first approach but those values are not used in the estimation.

In the current Mplus 8.5 implementation, missing values of predictors of Nominal/Count/Logit variables are allowed when the predictor is normally distributed. The estimation in the Nominal cases was discussed earlier. The estimation for the Count and Logit case is as follows. From the observed Count or Logit value  $Y$ , we generate the underlying variable  $Y^*$ , which is normally distributed. The missing predictor for  $Y$  is a missing predictor for  $Y^*$ . In that case, the conditional distribution for the missing predictor (conditional on  $Y^*$ ) is obtained from the multivariate normal distribution (for  $Y^*$  and the missing predictor) and is an explicit normal distribution.

## 2.5 Mixture Modeling

In Asparouhov and Muthén (2010a), three different Bayesian estimation methods are described for mixture models. The methods differ in how the latent class variable is updated: as a separate group, together with the latent continuous variables, or together with the underlying continuous variables when the model includes categorical variables. All three approaches extend naturally to the Nominal/Count/Logit variables because of the construction of the underlying latent variables described above. Not all methods are available for all models and not all methods are equally efficient. Mplus automatically selects the most efficient among the available methods, although it is possible for a specific method to be requested for a particular estimation via the ANALYSIS option CGENERATION.

## 2.6 Priors

In this section we discuss the choice of priors for the model parameters as well as the default settings in Mplus. For all regression parameters we use normal priors which are the conjugate priors for the above computation. The dispersion parameter prior is the inverse gamma prior which is also a conjugate prior. The default prior for the negative-binomial regression parameters is the improper and uninformative prior  $N(0, \infty)$ . For the logit and nominal regression parameters, the default prior is set to the weakly informative prior  $N(0, 5)$ . This weakly informative prior prevents estimation problems that sometimes occur when there is a categorical predictor and the contingency table of the dependent variable and the predictor contains empty cells. The weakly informative prior is also in line with the default priors used in Mplus for the probit regression parameters. The default prior for the dispersion parameter is  $IG(-1, 0)$ , which is an improper and uninformative prior with a constant density function over the interval  $(0, \infty)$ .

## 2.7 Model Fit

At this time model fit tools are fairly limited in Mplus 8.5. Significance of parameters can be evaluated through the credibility intervals and hypothesis involving multiple parameters can be evaluated via the Bayesian Wald test implemented in Mplus MODEL TEST, see Asparouhov and Muthén (2020). Model estimated distribution tables are also computed in Mplus and can be

obtained with the RESIDUAL option of the OUTPUT command. These distribution tables can be compared to the observed distribution tables to evaluate fit.

## 2.8 The General Model - SEM, Multilevel and Mixture Models

In this section, we described the general model implemented in the Bayesian framework of Mplus and show how Nominal/Count/Logit variables are incorporated in that model. We begin with the single level SEM model.

Let  $Y$  be a vector of  $p$  observed continuous dependent variables,  $\eta$  be a vector of  $m$  latent variables, and  $X$  be a vector of  $q$  independent observed variables. The structural equation model for these variables is given by the following equations

$$Y = \nu + \Lambda\eta + KX + \varepsilon \quad (36)$$

$$\eta = \alpha + B\eta + \Gamma X + \zeta. \quad (37)$$

To include categorical variables via the probit link function in the above model, we construct the underlying continuous variable as usual. For each categorical variable  $Y_j$  in the model, taking the values from 1 to  $k$ , we assume that there is a normally distributed latent variable  $Y_j^*$  and threshold parameters  $\tau_{1j}, \dots, \tau_{k-1j}$  such that

$$Y_j = t \Leftrightarrow \tau_{t-1j} \leq Y_j^* < \tau_{tj}, \quad (38)$$

where  $\tau_{0j} = -\infty$  and  $\tau_{kj} = \infty$ . This construction converts a categorical variable  $Y_j$  into an unobserved continuous variable  $Y_j^*$ . To include categorical variables in the above SEM model, we use  $Y_j^*$  in (36) instead of  $Y_j$ . For identification purposes the residual variance of  $Y_j^*$  is fixed to 1. This model represents the Bayesian SEM model implemented in Mplus Version 8.4.

Next we describe the inclusion of the Nominal/Count/Logit variables in the model. Suppose that  $N$  is a nominal variable taking values from 1, ...,  $k$ . The SEM model extend to the nominal variable as follows

$$P(N = j) = \frac{\text{Exp}(\nu_{N,j} + \beta_{N,j}Y + \Lambda_{N,j}\eta + \Gamma_{N,j}X)}{\sum_{i=1}^k \text{Exp}(\nu_{N,i} + \beta_{N,i}Y + \Lambda_{N,i}\eta + \Gamma_{N,i}X)}, \quad (39)$$

where for identification purposes the parameters  $\nu_{N,j}, \beta_{N,j}, \Lambda_{N,j}, \Gamma_{N,j}$  are fixed to 0 if  $j = k$ .



The model for a binary logistic link variable  $L$  is a special case of the nominal variable model ( $k = 2$ ) and reduces down to

$$P(L = 1) = \frac{1}{1 + \text{Exp}(\nu_L + \beta_L Y + \Lambda_L \eta + \Gamma_L X)}. \quad (40)$$

The general SEM model extends to a Negative-Binomial variable  $P$  as follows. The distribution of  $P$  is as in equation (9) where

$$p = \frac{1}{1 + \text{Exp}(\nu_P + \beta_P Y + \Lambda_P \eta + \Gamma_P X + \log(\alpha))} \quad (41)$$

if the variable is based on the NB-2 parameterization and

$$p = \frac{1}{1 + \text{Exp}(\nu_P + \beta_P Y + \Lambda_P \eta + \Gamma_P X)} \quad (42)$$

if the variable is based on the PG parameterization.

Note that in the above model the Nominal/Count/Logit variables do not have residual correlations the way continuous (observed and latent) and Probit-categorical variables do, i.e., through the covariances of  $\varepsilon$  and  $\zeta$ . The main way to model non-independence between such variables is through their common predictors. In particular, using common latent variables in the model leads to conditional non-independence the same way residual correlations do. Note also that the model as specified above does not allow for the Nominal/Count/Logit variables to be mediators, i.e., predictors for another variable in the model. This to some extent can also be remedied by specifying a latent continuous variable behind the variable and then the latent continuous variable can be used as a proxy mediator. Such an approach is often used with continuous observed variables where in addition, the residual variance is fixed to zero so that the continuous observed variable becomes identical to the latent variable. With Nominal/Count/Logit variables, we can not fix the "residual" to zero as such a residual does not exist. Therefore a latent variable proxy is equivalent to a latent factor measured by a single observed variable, which is usually an unidentified model. Further model restrictions must be devised to resolve this problem. Note also that the underlying continuous variable for Nominal/Count/Logit variables is not in the model. This is in contrast to the Probit-categorical variables where the underlying continuous variable can be used as a predictor for other variables. The underlying continuous variable for Nominal/Count/Logit variables is simply a tool that is used in the Bayesian estimation, but it bares no special meaning or interpretation in the actual SEM model.

### 2.8.1 Multilevel and Mixture Model Extensions

The extension of the SEM model to a two-level model is done as usual. Every regression or intercept parameter can be a cluster specific normally distributed random effect. On the between level, all random effects can be regressed on each other and other variables or be correlated with each other. The extension of the SEM model to a general mixture model is also done as usual. A latent categorical variable  $C$  is included in the model and every model parameter can be class-specific, i.e., the SEM model is different for every latent class. The latent class variable is treated as a nominal variable and can be regressed on other variables. If  $k$  is the number of classes in the model, the latent class distribution is given by

$$P(C = j|X) = \frac{\text{Exp}(\nu_{C,j} + \Gamma_{C,j}X)}{\sum_{i=1}^k \text{Exp}(\nu_{C,i} + \Gamma_{C,i}X)}. \quad (43)$$

Further discussion on the multilevel and mixture model extensions is available in Asparouhov and Muthén (2010a) and Asparouhov and Muthén (2019).

## 3 Examples

In this section we illustrate the Bayesian estimation with several simulation studies.

### 3.1 Logistic Latent Growth Model

A quadratic logistic growth model is given by the following equation

$$P(Y_t = 1) = \frac{1}{1 + \text{Exp}(i + st + qt^2)}, \quad (44)$$

where  $i$ ,  $s$  and  $q$  are normally distributed random effects

$$\begin{pmatrix} i \\ s \\ q \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \right). \quad (45)$$

We generate 8 equally spaced binary observations at times  $t$ , starting at -1.4 and ending at 1.4. In practical applications, the times of observations

Table 1: Quadratic logistic growth model: absolute bias(coverage)

Parameter	True Value	N=100	N=200	N=500	N=1000
$\mu_1$	-.3	.00(.95)	.00(.96)	.00(.99)	.00(.98)
$\mu_2$	.2	.01(.97)	.00(.99)	.00(.93)	.00(.95)
$\mu_3$	.25	.00(.93)	.00(.95)	.00(.94)	.00(.95)
$\sigma_{11}$	.5	.01(.98)	.00(.98)	.01(.94)	.00(.98)
$\sigma_{12}$	.1	.02(.92)	.00(.95)	.01(.93)	.01(.93)
$\sigma_{13}$	.05	.02(.98)	.02(.99)	.00(.98)	.00(.92)
$\sigma_{22}$	.3	.08(.96)	.05(.95)	.01(.95)	.00(.96)
$\sigma_{23}$	.05	.01(1.00)	.01(1.00)	.01(.95)	.00(.95)
$\sigma_{33}$	.2	.14(.96)	.10(.93)	.04(.94)	.02(.91)

are often set to integer values. With 8 observations the times can be set to 0,1,...,7 or 1,2,...,8. Using such alternative time scores yields equivalent models, however, there are certain implications that concern the Bayesian estimation. If the values of  $t$  are larger such as 7 or 8 (as compared to 1.4) the value of  $t^2$  will be as large as 49 or 64. This leads to a dramatic reduction in the scale of the quadratic term  $q$  and the variances of  $q$  could become very small. The implications of that are two-fold. First, the small variance of  $q$  may lead to slow convergence, worse mixing and larger model parameter autocorrelations across the MCMC iterations. Second, the quadratic effect may erroneously be deemed to be substantively minor. It is preferable to have the time scale be set so that the times of observations are approximately on a standard normal scale. With our choice of time scale, the mean of the time scores is 0 and the variance is 0.96, i.e., approximately standardized.

We conduct the simulation study using 100 replications for several different sample sizes  $N = 100, 200, 500, 1000$ . The results are presented in Table 1. Most parameters estimates appear to be unbiased and the coverage is near the nominal levels. For smaller sample sizes, however, the variance of the quadratic random effect has some bias. This bias occurs also with the ML estimator. For optimal performance, the model needs medium or higher sample sizes. For small sample sizes, the results should be interpreted cautiously. A more thorough simulation study would involve varying the number of time points in the growth model. Having fewer time points would likely

Table 2: Alternative estimation for quadratic logistic growth model: estimate (standard error)

Parameter	True Value	Logit Growth	Two-level Logit	Probit Growth
$\mu_1$	-.3	-.31(.03)	-.31(.03)	-.34(.03)
$\mu_2$	.2	.15(.02)	.15(.02)	.16(.03)
$\mu_3$	.25	.23(.03)	.23(.03)	.24(.03)
$\sigma_{11}$	.5	.48(.07)	.48(.07)	.60(.07)
$\sigma_{12}$	.1	.09(.03)	.09(.03)	.10(.04)
$\sigma_{13}$	.05	.08(.04)	.07(.04)	.04(.04)
$\sigma_{22}$	.3	.27(.04)	.28(.04)	.34(.04)
$\sigma_{23}$	.05	.01(.03)	.01(.04)	.01(.04)
$\sigma_{33}$	.2	.19(.05)	.20(.05)	.27(.05)

require even bigger sample sizes. Note also that the sample size requirements depend on the growth curve. A linear growth model would require less time points and smaller sample sizes than a quadratic curve model.

The above model can be estimated in two other completely different ways. First, the model can be estimated using the probit link function and then the mean parameter estimates can be multiplied by  $D = 1.749$  while the variance covariance estimates are multiplied by  $D^2$ . The value of  $D$  is chosen as in Camilli (2017). This approach provides an approximation to the logistic growth model and is based on the fact that the logit and the probit distribution functions are quite similar. In most applications, this approximation works sufficiently well. The advantage of the probit based model estimation is that it is faster when using Bayesian methods.

The second estimation alternative is the multilevel approach. The above model can be estimated as a two-level model where each cluster has 8 observations while  $t$  and  $t^2$  are treated as predictors for the binary variable. In this setup,  $s$  and  $q$  are simply the random coefficients for the predictors while  $i$  is the random intercept. In the Mplus language the variable  $i$  is also the between part of the binary variable.

To illustrate these alternative methods, we generate one data set using the above model with sample size  $N = 2000$  and we estimate the model using the three different approaches. The results are reported in Table 2. All three

approaches use Bayesian estimation and some of the small variation in the results can be attributed to randomization. We used 10000 MCMC iterations in this estimation to ensure that such variation is minimal. The logit growth model and the two-level logit model yield almost identical results as expected because the two models are the same. The probit growth model also yields similar results but some larger differences are visible in the estimates of the variances of the random effects.

### 3.2 Multilevel IRT

Multilevel IRT models have been discussed in Fox (2010), Asparouhov and Muthén (2016), and Muthén and Asparouhov (2018). With the Bayesian implementation of the logit link for binary variables, we can now estimate these models in the logit scale, instead of probit, which is traditionally used with the IRT models.

Consider the IRT model where  $P$  binary items measure a single latent variable. Observations are nested within groups, such as countries or regions, and all IRT model parameters are allowed to vary across the groups. This setup can be viewed as multiple-group IRT model or as multilevel IRT model. In multiple-group IRT models, the group specific parameters are treated as fixed, i.e., non-random parameters. In a multilevel IRT model, the parameters are treated as group specific random effects. Suppose that  $Y_{pij}$  is the  $p$ -th binary item for individual  $i$  in cluster or group  $j$ . The multilevel IRT model is given by the following equations

$$P(Y_{pij} = 1) = \frac{1}{1 + \text{Exp}(\tau_{pj} - \lambda_{pj}\eta_{ij})} \quad (46)$$

$$\eta_{ij} = \eta_{wij} + \eta_{bj} \quad (47)$$

$$\eta_{wij} \sim N(0, \text{Exp}(\psi_j)) \quad (48)$$

$$\psi_j \sim N(0, \psi_w) \quad (49)$$

$$\eta_{bj} \sim N(0, \psi_b) \quad (50)$$

$$\tau_{pj} \sim N(\tau_p, v_p) \quad (51)$$

$$\lambda_{pj} \sim N(\lambda_p, w_p). \quad (52)$$

The difficulty parameter  $\tau_{pj}$  varies across clusters and its mean across all the groups is the average difficulty parameter  $\tau_p$ . The discrimination parameter

Table 3: Two-level IRT

Parameter	Absolute bias(Coverage)
$\tau_1$	.01(.92)
$\lambda_1$	.03(.92)
$v_1$	.01(.96)
$w_1$	.00(.95)
$\psi_w$	.01(.93)
$\psi_b$	.00(.95)

$\lambda_{pj}$  varies across clusters and its mean across all the groups is the average difficulty discrimination  $\lambda_p$ . The factor mean in group  $j$  is  $\eta_{bj}$ , i.e., we can estimate a group specific factor mean. Also  $\eta_{wij} = \eta_{ij} - \eta_{bj}$  is the individual-specific factor deviation from the group specific mean  $\eta_{bj}$ . The factor variance in group  $j$  is  $Exp(\psi_j)$ , i.e., we can estimate a group specific factor variance through the random effects  $\psi_j = \log(Var(\eta_{wij}|j))$ . In a standard IRT model, the factor mean is fixed to 0 and the factor variance is fixed to 1 for identification purposes. These identification restrictions are now replaced by fixing the mean of the random effect  $\eta_{bj}$  to 0 and by fixing the mean of the random effect  $\psi_j$  to 0, respectively. The above model is able to estimate group specific factor mean and variance without having to assume metric or scalar measurement invariance. The model has  $4P + 2$  parameters:  $\tau_p$ ,  $\lambda_p$ ,  $v_p$ ,  $w_p$ , for  $p = 1, \dots, P$  as well as  $\psi_w$  and  $\psi_b$ .

We conduct a simulation study to illustrate the quality of the Bayesian estimation using the following parameter values  $\tau_p = .3$ ,  $\lambda_p = 1.4$ ,  $v_p = w_p = \psi_w = \psi_b = 0.1$ . Using  $P = 8$  indicators, we generate 100 samples with 200 groups of size 30. The results of the simulation study for some of the parameters are reported in Table 3 and show minimal bias and coverage near the nominal level.

In practical applications, the above model should be followed up by further analysis. Random effects with insignificant variance estimates should be eliminated from the model. If the credibility interval of the variance of a random effect is very close to zero, it should be considered insignificant and the random effect should be replaced by a non-random group invariant parameter. The reduced model flexibility will actually improve the accuracy for all other model parameters.

It is important to note here that all the parameters in the above model are between level parameters. This implies that the sample size used in the estimation of the parameters is the number of groups. With very few groups, this model would not be feasible. For example, with less than 30 groups, the model would be difficult to estimate and even if it is possible to estimate it, the random effect variance estimates would depend on the priors. In addition, the standard errors are likely to be large and most if not all random effect variances would not be significant. When the number of groups is small, the alignment methodology described in Asparouhov and Muthén (2014b) can accomplish the same multiple group flexibility without relying on random effects.

Alternatively, with a small number of groups, one can pursue somewhat less flexible multilevel IRT models along the line of traditional multilevel SEM models. One such formulation is as follows

$$P(Y_{pij} = 1) = \frac{1}{1 + \text{Exp}(\tau_{pj} - \lambda_{wp}\eta_{wij} - \lambda_{bp}\eta_{bj})} \quad (53)$$

$$\eta_{wij} \sim N(0, 1) \quad (54)$$

$$\eta_{bj} \sim N(0, 1) \quad (55)$$

$$\tau_{pj} \sim N(\tau_p, v_p). \quad (56)$$

This model does not include random loadings and random factor variances. The between factor  $\eta_{bj}$  can be interpreted as a cluster specific mean but only if the between and within level loadings are proportional. More specifically, if  $\lambda_{bp} = s\lambda_{wp}$  for a proportionality scale parameter  $s$ , then  $\lambda_{wp}\eta_{wij} + \lambda_{bp}\eta_{bj} = \lambda_{wp}(\eta_{wij} + s\eta_{bj})$ . The last expression allows us to interpret  $s\eta_{bj}$  as the cluster specific factor mean. The proportionality of the loadings can be tested with MODEL TEST in Mplus. If the proportionality is rejected, allowing for different within and between level loadings becomes an advantage over the multilevel IRT model given in (46-52).

Next we conduct a simulation study to evaluate the performance of model (53-56) for small samples. We use the following parameter values for this simulation study  $\lambda_{wp} = 1.4$ ,  $\lambda_{bp} = 0.5$ ,  $\tau_p = 0.3$  and  $v_p = 0.1$ . Again using  $P = 8$ , we generate 100 samples for several combinations of number of groups and group sizes. In this simulation study, we also illustrate the effect of weakly informative priors for the random effect variance parameters  $v_p$ . The Mplus default prior is the improper prior with density function of 1 on the

Table 4: Two-level IRT, multilevel SEM style: Absolute bias(Coverage)

Groups	40	40	40	40	80	200
Group sizes	15	15	15	30	15	15
$v_p$ priors	-	IG(3,0.4)	IG(3,1)	-	-	-
$\tau_1$	.01(.98)	.00(.96)	.01(.99)	.02(.98)	.00(.95)	.00(.96)
$\lambda_{w1}$	.09(.93)	.08(.93)	.10(.91)	.02(.97)	.06(.94)	.02(.94)
$\lambda_{b1}$	.01(.98)	.03(.94)	.01(.96)	.04(.98)	.02(.93)	.01(.96)
$v_1$	.10(.94)	.04(1.00)	.15(.07)	.04(.96)	.05(.92)	.02(.92)

interval from 0 to infinity, which is specified as  $IG(-1, 0)$ . As an alternative, we also estimate the model using two different weakly informative priors for  $v_p$ . The first prior is  $IG(3, .4)$ , which has a mode at the true value of .1 and standard deviation of 0.2. The second prior is  $IG(3,1)$ , which has a mode at 0.25 and standard deviation of 0.5.

The results of the simulation study for some of the parameters are reported in Table 4. Some bias can be found in the estimates and the bias generally appears to decrease depending on how the sample size is increased. If we increase the groups sizes, the within level loadings improve. If we increase the number of groups, the between level loadings improve. If the number of group increases, the bias will disappear even if the group sizes stay small. Such a result is in line with asymptotic theory, which guarantees unbiased model estimates as the number of groups increases, regardless of the size of the groups.

The weakly informative prior has a fairly substantial effect on the random effect variance parameter estimates for the sample with 40 groups. The results illustrate that the estimates can be improved with such priors. However, if the prior is not set well, the bias can actually become worse. This is exactly what happened with the  $IG(3, 1)$  prior. The effect of the prior is intuitively easy to understand. If the non-informative prior yields a posterior distribution with standard deviation of 0.2 and we add a weakly informative prior with the same standard deviation, the estimate of the variance would then become approximately the average of the non-informative prior estimate and the median of the weakly informative prior. The underlying issue here is that little information is extracted from the data regarding the variance of the random threshold. Large standard errors for these parameters, relative



Table 5: Simple Two-level IRT: Absolute bias(Coverage)

Groups	40	20	10
Group sizes	15	20	30
$\tau_1$	.02(.93)	.01(.93)	.03(.98)
$\lambda_1$	.03(.92)	.06(.92)	.09(.91)
$v$	.02(.98)	.04(.93)	.10(.90)

to the size of the variance, essentially results in non-significant variance parameters. In fact, in the particular setting of the first column of Table 4, if we fix the  $v_p$  parameters to zero, we get the same estimates for the remaining parameters. This also confirms that at this sample size and level of threshold non-invariance, the random effects  $\tau_{pj}$  are not essential.

Next, we consider an even simpler multilevel IRT model. In this model the thresholds and the loadings are group invariant as well as the factor variance. Only the factor mean is allowed to vary across groups. The model is described as follows

$$P(Y_{pij} = 1) = \frac{1}{1 + \text{Exp}(\tau_p - \lambda_p \eta_{ij})} \quad (57)$$

$$\eta_{ij} = \eta_{wij} + \eta_{bj} \quad (58)$$

$$\eta_{wij} \sim N(0, 1) \quad (59)$$

$$\eta_{bj} \sim N(0, v). \quad (60)$$

Because of its simplicity, the model can be estimated with very small samples fairly well. Using the earlier setup with  $P = 8$ ,  $\tau_p = 0.3$ ,  $\lambda_p = 1.4$  and  $v = 0.2$ , we conduct a simulation study with varying number of groups and group sizes across 100 replications. The results for some of the parameters are reported in Table 5.

### 3.3 Multilevel autoregressive model for count data

Polson et al. (2013) describe an autoregressive model for count data. The model is based on single time series data using the PG Negative-Binomial parameterization. Here we describe a multilevel version for that model where the data consists of multiple time-series count data for a group of individuals

Table 6: Multilevel autoregressive model for count data, N=500, T=10

Parameter	True Value	Absolute bias(Coverage)
$\nu$	.3	.01(.92)
$\rho$	.5	.01(.87)
$\alpha$	.4	.02(.88)
$v_1$	.2	.00(.93)
$v_2$	.3	.02(.87)

observed across time. We use the NB-2 and Poisson parameterizations for this illustration.

Suppose that  $Y_{it}$  is a count observation for individual  $i$  at time  $t$ . We consider the following autoregressive model

$$Y_{it} \sim NB2(\nu_i + \eta_{it}, \alpha) \quad (61)$$

$$\nu_i \sim N(\nu, v_1) \quad (62)$$

$$\eta_{it} = \rho\eta_{i,t-1} + \varepsilon_{it}, \text{ for } t = 2, \dots, T \quad (63)$$

$$\varepsilon_{it} \sim N(0, v_2). \quad (64)$$

The model has 5 parameters:  $\nu$ ,  $\rho$ ,  $\alpha$ ,  $v_1$  and  $v_2$ . We generate 100 data sets with  $N = 500$  individuals and  $T = 10$  observations. The results of the simulation study are reported in Table 6 and show that the Bayesian estimation performs well.

The autoregressive coefficient in this model is time invariant. It is possible to estimate a model with time-specific autocorrelation coefficients, however, such a model is somewhat poorly identified and it is not recommended. The model is estimated in Mplus as a single level multivariate model. Full DSEM model flexibility as in Asparouhov et al. (2017) is currently not implemented. This has two implications. First, models with  $T > 50$  are going to be very slow to estimate. Second, the autocorrelation coefficient  $\rho$  can not be subject specific. The above model also does not perform very well with smaller values of  $T$ . Using  $T = 5$  for example yields larger biases and lower coverage.

In practical applications, especially when  $T$  is not large, it is recommended that an additional parameter is estimated for the variance of the initial starting value  $\eta_{i,1}$ . Assuming a stationary time-series model for  $\eta_{i,t}$ ,

Table 7: Multilevel autoregressive model for count data, N=100, T=5: Absolute bias(Coverage)

Parameter	True Value	Poisson	Negative-Binomial
$\nu$	.3	.00(.92)	.14(.58)
$\rho$	.5	.06(.94)	.26(.65)
$\alpha$	.4	-	.31(.44)
$v_1$	.2	.02(.88)	.03(.93)
$v_2$	.3	.02(.90)	.28(.50)
convergence rate		100%	72%

the variance  $Var(\eta_{i,t}) = v_2/(1 - \rho^2)$ . At the first time point, however,  $\eta_{i,1}$  is not predicted by another variable and thus the residual variance for  $\varepsilon_{i1}$  should be set to  $v_2/(1 - \rho^2)$ . This would involve a complex parameter constraint, however. A simpler alternative is to just estimate the first residual variance as a separate model parameter, i.e., constrain the residual variance of  $\eta_{i,t}$  to be time-invariant only for times  $2, \dots, T$ .

The above model performs better with the Poisson distribution instead of the Negative-Binomial distribution, particularly for smaller sample sizes. That is because the autoregressive model is imposed on the latent variables and the variances of the latent variables are somewhat confounded with the dispersion parameters. Recall that a Poisson distribution where the mean parameter is a Gamma distributed random effect is a Negative-Binomial distribution, i.e., a random effect essentially introduces and models over-dispersion. In the auto-regressive model, the latent variables are primarily identified through the across-time correlations between the variables, i.e., the autoregressive process is truly identified by the across time correlations rather than by fitting the over-dispersion in the variables. When the time series is short ( $T$  is small), however, the autoregressive information is weak (due to small  $T$  the model has difficulty in distinguishing between correlation that is due to observations nested within person and due to observations taken in proximity of time) and could become conflated with the over-dispersion.

To illustrate this point we conduct a simulation study with  $T = 5$  and  $N = 100$  for auto-regressive Poisson model and auto-regressive Negative-Binomial model. The auto-regressive Poisson model is described as the auto-

regressive Negative-Binomial model given above with  $\alpha = 0$ , i.e.,

$$Y_{it} \sim \text{Pois}(\nu_i + \eta_{it}) \quad (65)$$

$$\nu_i \sim N(\nu, v_1) \quad (66)$$

$$\eta_{it} = \rho\eta_{i,t-1} + \varepsilon_{it}, \text{ for } t = 2, \dots, T \quad (67)$$

$$\varepsilon_{it} \sim N(0, v_2). \quad (68)$$

We generate 100 replications for each of the two distribution types and analyze the data with the same distribution type. The results of the simulation study are reported in Table 7. The Bayesian estimation for the Poisson autoregressive model performs well, while for the Negative-Binomial it does not. For small  $T$ , the Poisson auto-regressive model is a well identified model that can be used in practice. The Negative-Binomial model for small  $T$  is somewhat poorly identified and can not be recommended for practical applications. In the Negative-Binomial model, the dispersion parameter estimates are close to 0. This implies that the Poisson autoregressive model provides sufficient fit for the data even when it is generated by the Negative-Binomial autoregressive model when  $T$  is small.

The above model can be incorporated into a cross-lagged panel model (RI-CLPM) as in Hamaker et al. (2015). In particular, having a second time series, for example with continuous items, could actually improve the identifiability of the model. That is because the cross-lagged relations can contribute to the measurement of the latent factors  $\eta_{it}$  used in the count time series.

### 3.4 Multilevel Nominal regression

In this section we consider the nominal regression model in two-level settings. In particular, we illustrate the concepts of latent centering for the covariates, see Asparouhov and Muthén (2019), as well as the random regression coefficients and intercepts.

Let  $N_{ij}$  be a nominal variable for individual  $i$  in cluster  $j$  and  $X_{ij}$  be the corresponding predictor. Let  $K$  be the number of unordered categories for the nominal variable. The two-level nominal regression model is given by the following equations

$$X_{ij} = X_{w,ij} + X_{b,j} \quad (69)$$

$$X_{w,ij} \sim N(0, \psi_w) \quad (70)$$

Table 8: Two-level Nominal regression

Parameter	True Value	Absolute bias(Coverage)
$\alpha_1$	1	.01(.95)
$\alpha_2$	.5	.00(.95)
$\beta_1$	.4	.01(.95)
$\beta_2$	-.4	.01(.96)
$\gamma_1$	.6	.01(.96)
$\gamma_2$	.2	.01(.96)
$\theta_1$	.5	.05(.94)
$\theta_2$	.5	.05(.92)
$\sigma_1$	.2	.02(.95)
$\sigma_2$	.2	.01(.97)
$\rho$	.2	.03(.97)
$\psi_w$	1	.00(.96)
$\psi_b$	.3	.03(.86)
$\mu$	0	.00(.96)

$$X_{b,j} \sim N(\mu, \psi_b) \quad (71)$$

$$P(Y_{ij} = k) = \frac{\text{Exp}(\alpha_{jk} + \beta_{jk}X_{w,ij} + \gamma_k X_{b,j})}{\sum_{i=1}^K \text{Exp}(\alpha_{jk} + \beta_{jk}X_{w,ij} + \gamma_k X_{b,j})} \quad (72)$$

$$\alpha_{jk} \sim N(\alpha_k, \theta_k) \quad (73)$$

$$\beta_{jk} \sim N(\beta_k, \sigma_k). \quad (74)$$

The parameters  $\alpha_{jK}$ ,  $\beta_{jK}$  and  $\gamma_K$  are all zero for identification purposes. In this model,  $X_{b,j}$  is the mean of  $X_{ij}$  in cluster  $j$ . This mean is an unobserved latent variable. In principle, it is possible to use the average of  $X_{ij}$  in cluster  $j$  as the mean. However, it has been shown that such an approach yields biased estimates due to not accounting for the measurement error in that average, see Lüdtke et al. (2008), Asparouhov and Muthén (2006) and Asparouhov and Muthén (2019). The variable  $X_{w,ij}$  is the group centered covariate. In the above model, the effect of the covariate has two separate effects: the effect of the cluster mean  $X_{b,j}$  as well as the effect of the group centered covariate  $X_{w,ij}$ . The difference between these two effects is called the contextual effect, see Lüdtke et al. (2008). If the covariate  $X_{ij}$  is not

Table 9: Two-level Nominal regression with non random slopes: Absolute bias(Coverage)

Parameter	True Value	$C = 100, L = 15$	$C = 200, L = 20$
$\alpha_1$	1	.02(.98)	.01(.93)
$\alpha_2$	.5	.02(.94)	.00(.95)
$\beta_1$	.4	.01(.91)	.00(.91)
$\beta_2$	-.4	.00(.95)	.00(.96)
$\gamma_1$	.6	.01(.93)	.04 (.96)
$\gamma_2$	.2	.02(.97)	.00(.92)
$\theta_1$	.5	.10(.96)	.03(.91)
$\theta_2$	.5	.06(.96)	.03(.94)
$\rho$	.2	.04(.92)	.02(.93)
$\psi_w$	1	.00(.95)	.01(.95)
$\psi_b$	.3	.01(.91)	.01(.90)
$\mu$	0	.01(.94)	.01(.95)

decomposed as the centered portion and the centering portion, and is used directly in the nominal regression, we would estimate an “uninterpretable blend” between the two different effects, see Raudenbush and Bryk (2002). Since  $X_{w,ij}$  varies within cluster, we can estimate a cluster specific random effect  $\beta_{jk}$ . Similarly, the intercept in the nominal regression can be estimated as a random effect  $\alpha_{jk}$ . The effect of  $X_{b,j}$  can not be cluster specific. The model has a total of  $2K - 1$  random effects:  $\alpha_{jk}$ ,  $\beta_{jk}$  and  $X_{b,j}$ . Correlations between these random effects can be added to the above model with one exception. The correlation between  $\alpha_{jk}$  and  $X_{b,j}$  would not be identified because it is confounded with the parameters  $\gamma_k$ . If we want to estimate the correlation between these two random effects we must remove  $\gamma_k$  from the model, otherwise the model would not be identified.

We conduct a simulation study to evaluate the performance of the Bayesian estimation for the above model. Using  $K = 3$ , we generate 100 samples with 200 groups of size 30. In this simulation study we include the parameter  $\rho = Cov(\alpha_{1j}, \alpha_{2j})$ . The results, reported in Table 8, show that the bias is minimal and the coverage is near the nominal levels.

Model (69-74) has a large number of random effects and is expected to

require a fairly large sample for optimal performance. For smaller samples, simpler two-level nominal regression should be explored. One such model is the model where the random slope variance  $\sigma_k$  is fixed to 0, i.e., the random effect for  $X_{w,ij}$  is replaced by a non-random regression. We conduct a simulation study to evaluate the performance of the Bayesian estimation for that model using different number of clusters  $C$  and cluster sizes  $L$ . The results of this simulation are reported in Table 9. The bias is minimal and the coverage is near the nominal levels. Slightly larger bias is visible for the variance of the random intercept when the number of clusters is 100. As for other multilevel models with smaller number of clusters, weakly informative priors can be used to reduce that bias.

### 3.5 Nominal factor analysis

Revuelta et al. (2020) discusses factor analysis models where all the factor indicators are nominal variable. Such models can be estimated with the ML estimator using numerical integration as long as the number of factors is not large. Models with up to three factors can be estimated fairly well with the ML estimator. However, for models with more than three factors, the ML estimation will be very slow and will be prone to convergence problems. The Bayesian estimator can be used as an alternative to the ML estimator, because it will not be limited by the number of factors. In this section we evaluate the performance of the Bayesian estimator for such models and provide some insights into this fairly novel modeling technique.

Using nominal factor indicators, instead of ordered categorical, is necessary for those situations when the outcome can not be regarded as an ordered variable. Nominal variables are also a useful modeling alternative when the proportional odds ratio, assumed with the ordered logistic model, does not hold. The nominal model, however, has some drawbacks as well and it should not be used routinely instead of ordered models. One drawback is that the nominal model is less parsimonious. Many more parameters will be estimated with the nominal model, and very likely, some loss of power will occur. A nominal model will likely require bigger sample size than an ordered logistic model. Currently, there is no simple statistical tool that can be used to determine whether a variable should be treated as ordered or as nominal. This is particularly the case in factor analysis where the predictors is not observed and the proportionality of the odds ratio can not be easily evaluated. Potentially, one can lean into the substantive interpretation of the indicator

and assume that the substantive interpretation is the correct choice. Such an approach, however, has some limitations as well. First, the substantive interpretation can be ambiguous and both nominal and ordered may be viable options. Second, the substantive interpretation does not necessarily have to match with the best statistical model. This issue is particularly acute for small samples, where lack of power is bound to interfere with an attempt to make a rigorous choice. The Bayesian estimation of the nominal factor analysis model has one advantage over the ML estimation in this regard. Plausible values can be imputed for the factor, which can subsequently be used to study the nominal v.s. ordinal nature of the variable in separate analysis.

Another practical aspect in the nominal v.s. ordinal dilemma is the fact that the categories can be ordered in  $L!$  different ways where  $L$  is the number of categories. This means that the comparison between the nominal model and ordered model is essentially a comparison between the nominal model and  $L!$  ordered models. If the latent factor is imputed, log-likelihood comparison could potentially be used to guide an informed choice. Note, however, that if there is only one predictor for the categorical variable, the best ordering for the categories of the nominal variable is easy to obtain. Suppose that we have a nominal variable  $N$  regressed on a covariate  $X$  and the regression coefficients are  $\beta_1, \dots, \beta_{L-1}, \beta_L = 0$ . To find the best ordering of the categories most suitable for an ordered logistic regression of  $N$  on  $X$ , we have to order the regression coefficients  $\beta_i$  so that they appear monotonic. Increasing or decreasing order works equally well. This is because in a logistic regression with a positive coefficient, higher values of the predictor implies an increase in the likelihood for the higher values of the categorical variable. In the nominal regression, such a relationship between the categorical variable and the covariate exist only when the regression coefficients are monotonic. To be more specific, if a data is generated from a logistic regression and is then analyzed with a nominal regression, the regression coefficients will appear in a monotonic order. Conversely, if we have a categorical variable that is treated as nominal, and we consider the question of whether or not the variable can be treated as ordered variable, the best fitting model would be obtained when the categories are ordered so that the nominal regression coefficients are monotonic.

In the presence of multiple covariates, if the same category ordering yields monotonic regression coefficients for every covariate, then clearly that ordering would be optimal. However, it is certainly possible that the nominal



regression coefficients of one covariate yields ordering that is different from the ordering implied by another covariate. This could be a clear sign that ordered regression is not appropriate. However, the ordering of the coefficients is subject to these coefficients standard errors and the mismatch in the ordering of the coefficients may not be statistically significant. In such situations, comparing the log-likelihoods among the  $L!$  different ordered logistic regressions may be necessary.

On the opposite spectrum of this discussion is the proportional odds assumption underlying the ordered logistic model. Suppose that a categorical variable  $U$  is a measurement indicator for a factor  $\eta$  through an ordered logistic regression. The model implies that

$$\log(P(U \leq j)/P(U > j)) = \alpha_j + \lambda\eta. \quad (75)$$

The proportional odds assumption refers to the fact that the coefficient  $\lambda$  is independent of  $j$ , i.e, it is the same across all  $j = 1, \dots, L - 1$ , see Agresti (1990). If this assumption does not hold, the ordered logistic measurement model should be replaced with a nominal measurement model.

Here we describe three different ways to test the proportional odds assumption for a latent variable  $\eta$ . The first method is as follows. Using the ML estimator, the factor analysis model is estimated as well as the factor scores. At this point, we can simply use Brant's (1990) test for the proportional odds assumption. This will be obtained in Mplus by estimating the logistic regression of  $U$  on the factor score with the ML estimator. The drawback of this method is that the uncertainty/measurement error in the factor score is not accounted for, which can result in underestimation of the p-value in Brant's test. It would be preferable, to impute the factor, and then use Brant's test using the plausible values. However, currently Brant's test is not available in Mplus for multiple imputed data.

The second and the third method we describe here can be used with multiple imputed data. To obtain plausible values for the factor, the factor model must be estimated with the Bayesian estimator, using the probit link function for all ordered indicators. Currently, the logit link function is not available in the Mplus Bayesian framework for non-binary variables and thus we must switch to the probit link function for those indicators. It is unlikely that such a switch will create estimation issues for the purpose of testing the proportional odds assumption. Given the plausible values for the factor, the second method proceeds as follows. We define  $L - 1$  binary variables  $W_j$  for

$j = 1, \dots, L - 1$  as follows

$$W_j = \begin{cases} 0, & \text{if } U \leq j \\ 1, & \text{if } U > j. \end{cases} \quad (76)$$

The next step involves simultaneously estimating the  $L - 1$  logistic regressions of  $W_j$  on the plausible values  $\eta$ , using the multiple imputed data for  $\eta$  and the ML estimator. In this estimation, we include Wald's test for the hypothesis that all  $L - 1$  regression coefficients are equal. For the validity of Wald's test for multiple imputed data see Asparouhov and Muthén (2010b). The Wald's test in Mplus is specified using the MODEL TEST command and is our second method for testing the proportional odds assumption.

The third method also uses the multiple imputed data for the factor and Wald's test. For this method, we estimate the following non-proportional odds ratio logistic regression model

$$P(U = j) = \frac{1}{1 + \text{Exp}(-\tau_j + \beta\eta)} - \frac{1}{1 + \text{Exp}(-\tau_{j-1} + \beta\eta)}, \quad (77)$$

where as usual  $\tau_0 = -\infty$ ,  $\tau_L = \infty$ . To make this model into a non-proportional odds ratio model, we introduce the following additional model constraints. For  $j = 2, \dots, L - 1$

$$\tau_j = \alpha_j + \beta_j\eta. \quad (78)$$

In the ML estimation of Mplus, such model constraints are introduced using the CONSTRAINT option of the VARIABLE command where  $\eta$  is specified. In addition, the parameters  $\alpha_j$  and  $\beta_j$  are declared as NEW parameters in the MODEL CONSTRAINT command, where also equations (78) are specified. Finally, the proportional odds ratio test is obtained using Wald's test on the hypothesis that  $\beta_j = 0$  for  $j = 2, \dots, L - 1$ .

Next we illustrate the Bayesian estimation of the nominal factor analysis model with a simulation study. The model we consider has 4 nominal variables  $N_1, \dots, N_4$  and a covariate  $X$ . The latent factor  $\eta$  is measured by the 4 indicators and is predicted by the covariate. Let  $L_i$  denote the number of categories for the nominal variable  $N_i$ . In this example, we set  $L_1 = L_2 = 3$ ,  $L_3 = 2$ , and  $L_4 = 4$ . The model is described by following equations

$$P(N_i = j) = \frac{\text{Exp}(\alpha_{ij} + \lambda_{ij}\eta)}{\sum_{j=1}^{L_i} \text{Exp}(\alpha_{ij} + \lambda_{ij}\eta)} \quad (79)$$

Table 10: Nominal Factor Analysis with 4 indicators: Absolute bias / Coverage / MSE

Parameter	N=1000	N=2000	N=1000 + Z
$\alpha_{11}$	.01/.93/.01	.00/.94/.00	.00/.96/.01
$\alpha_{21}$	.01/.94/.01	.01/.90/.00	.00/.93/.01
$\alpha_{31}$	.00/.93/.01	.00/.95/.00	.01/.93/.01
$\alpha_{41}$	.00/.96/.02	.00/.93/.01	.01/.91/.01
$\lambda_{11}$	.06/.92/.22	.01/.95/.03	.03/.91/.02
$\lambda_{21}$	.00/.94/.02	.01/.93/.01	.00/.97/.01
$\lambda_{31}$	.02/.94/.01	.00/.98/.01	.01/.94/.01
$\lambda_{41}$	.08/.89/.15	.02/.91/.03	.00/.94/.02
$\beta$	.03/.91/.01	.02/.93/.00	.00/.98/.00

$$\eta = \beta X + \varepsilon. \quad (80)$$

For identification purposes, the intercept in (80) is fixed to zero and  $\varepsilon$  is assumed to have a standard normal distribution, i.e., the residual variance of the factor is fixed to 1. In addition, both  $\alpha_{ij}$  and  $\lambda_{ij}$  are fixed to 0 when  $j = L_i$ . We generate 100 data sets with sample size  $N=1000$  and  $N=2000$  and the following parameter values  $\alpha_{11} = .5$ ,  $\alpha_{12} = -.5$ ,  $\alpha_{21} = .7$ ,  $\alpha_{22} = 0$ ,  $\alpha_{31} = -1$ ,  $\alpha_{41} = .4$ ,  $\alpha_{42} = -.2$ ,  $\alpha_{43} = .2$ ,  $\lambda_{11} = 1$ ,  $\lambda_{12} = .7$ ,  $\lambda_{21} = .3$ ,  $\lambda_{22} = .5$ ,  $\lambda_{31} = -.5$ ,  $\lambda_{41} = .6$ ,  $\lambda_{42} = .3$ ,  $\lambda_{43} = -.5$ ,  $\beta = .4$ . The results for a subset of the parameters are reported in Table 10. Some small biases are visible for some of the loadings parameter in the  $N = 1000$  case and some rather large MSE values can be seen for those parameters. Increasing the sample size to  $N = 2000$  appears to resolve both. The reduction in the MSE is dramatic. This indicates that the Bayesian estimation of the nominal factor analysis model may need larger samples for optimal performance. In fact, the ML estimator appear to perform better for smaller sample sizes. With the alternative factor analysis parameterization, where a loading is fixed and the factor variance is estimated, the Bayesian estimation appears to perform even worse in terms of convergence rates, quality of mixing, as well as bias and MSE.

For comparison purposes, we also conduct the following simulation study. To the above model with 4 nominal indicators, we add one additional con-

Table 11: Nominal Factor Analysis with 10 indicators: Absolute bias(Coverage)

Parameter	N=1000	N=500	N=300
$\alpha_{11}$	.00(.96)	.00(.96)	.02(.95)
$\alpha_{12}$	.00(.97)	.00(.94)	.00(.93)
$\lambda_{11}$	.02(.91)	.03(.96)	.07(.89)
$\lambda_{12}$	.02(.90)	.01(.89)	.01(.92)

tinuous factor indicator  $Z$ . We set the mean of the indicator to 0, the factor loading to 1 and the residual variance to 1. We generate and analyze 100 data sets with  $N = 1000$  observations. The results of this simulation study, also shown in Table 10, indicate that the biases in the parameters are resolved as well as the MSE. The additional continuous factor indicator appears to have stabilized the Bayesian estimation. This simulation study also suggests that a factor analysis that includes not just nominal indicators, but a mixture of different types of indicators, may be the most practical modeling approach, particularly when the sample sizes are small or moderate. In fact, the estimation of the nominal factor analysis can be improved by also adding ordinal indicators, i.e., indicators that help with measuring the latent factor but are less demanding than nominal variables in terms of the number of additional parameters that are to be estimated. This further emphasizes the need to improve our understanding of when a categorical variable should be specified as nominal or as ordered variable. The estimation is also improved by the presence of factor predictors, which in a way also improve the uncertainty in the measurement model.

The model estimation can also be improved by simply adding more nominal indicators. To illustrate this, we conduct a simulation study with 10 nominal indicators each with 3 categories, measuring 1 latent factor (and no covariates). We generate and analyze 100 samples with  $N = 1000, 500$  and 300 observations, using the following parameter values:  $\lambda_{i1} = 1, \lambda_{i2} = 0.5$  and  $\alpha_{ij} = 0$ . The results for a subset of the parameters are reported in Table 11. The bias is minimal and the coverage is near the nominal levels. Note, however, that in practical situations, some of the nominal indicators may carry very little information. If some of the categories are rare, the data will contain very little information about those category-specific parameters.

In turn, those parameters will have large posterior distributions that likely will lead to slow convergence. To avoid such convergence issues it may be necessary to add stronger weakly informative priors for the problematic parameters. For example, the Mplus default prior of  $N(0, 5)$  could be replaced by a prior of  $N(0, 1)$ . Having indicators with rare categories will naturally lead to poor estimation for these category-specific parameters, however, this will not compromise the estimation of the rest of the model parameters.

## References

- [1] Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.
- [2] Asparouhov, T., & Muthén, B. (2006). Constructing covariates in multilevel regression. *Mplus Web Notes*: No. 11.  
<http://www.statmodel.com/download/webnotes/webnote11.pdf>
- [3] Asparouhov, T. & Muthén, B. (2010a). Bayesian analysis using Mplus: Technical implementation. Technical appendix.  
[www.statmodel.com/download/Bayes3.pdf](http://www.statmodel.com/download/Bayes3.pdf)
- [4] Asparouhov, T. & Muthén, B. (2010b). Chi-square statistics with multiple imputation (Mplus technical appendices: Version 2). Los Angeles, CA: Muthén & Muthén. <http://www.statmodel.com/download/MI7.pdf>
- [5] Asparouhov, T. & Muthén, B. (2014a). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 329-341.
- [6] Asparouhov, T. & Muthén, B., (2014b). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 495–508. <https://doi.org/10.1080/10705511.2014.919210>
- [7] Asparouhov, T., & Muthén, B. (2016). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in realworld applications* (pp. 163–192). Charlotte, NC: Information Age.
- [8] Asparouhov, T., Hamaker, E., & Muthén, B. (2017). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 359–388. doi: 10.1080/10705511.2017.1406803
- [9] Asparouhov, T. & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, 26, 119-142.

- [10] Asparouhov, T. & Muthén, B. (2020). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*. doi.org/10.1080/10705511.2020.1764360
- [11] Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*, 46(4), 1171–1178. doi:10.2307/2532457
- [12] Camilli, G. (2017) The Scaling Constant D in Item Response Theory. *Open Journal of Statistics*, 7, 780-785. doi: 10.4236/ojs.2017.75055.
- [13] Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, 20(1), 102-116. DOI: 10.1037/a0038889
- [14] Hilbe, J.M. (2011). *Negative binomial regression*. Second edition. New York: Cambridge University Press.
- [15] Fox, J. P. (2010). *Bayesian item response theory*. New York, NY: Springer
- [16] Lüdtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203-229.
- [17] Muthén, B., & Asparouhov, T. (2007). Growth mixture analysis: Models with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (eds.), *Advances in Longitudinal Data Analysis*. Chapman & Hall/CRC Press.
- [18] Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups. *Sociological Methods & Research*, 47, 637– 664. doi:doi:10.1177/0049124117701488
- [19] Neelon B. (2019). Bayesian Zero-Inflated Negative Binomial Regression Based on Polya Gamma Mixtures. *Bayesian Analysis*, 14(3), 829-55.
- [20] Pillow, J. and Scott, J. (2012). Fully Bayesian inference for neural models with negative-binomial spiking. In Bartlett, P., Pereira, F., Burges,

C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 25*, 1907–1915. MIT Press.

- [21] Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108, 1339–1349.
- [22] Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- [23] Revuelta J., Maydeu-Olivares A. & Ximénez C. (2020) Factor Analysis for Nominal (First Choice) Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 27, 781-797, DOI: 10.1080/10705511.2019.1668276
- [24] Windle, J., Polson, N. G., and Scott, J. G. (2014). Sampling Polya-Gamma random variates: alternate and approximate techniques. arXiv preprint arXiv:1405.0506. <https://arxiv.org/abs/1405.0506>
- [25] Zhou, M. and Carin, L. (2015). Negative Binomial Process Count and Mixture Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 307–320.