General Random Effect Latent Variable Modeling: Random Subjects, Items, Contexts, and Parameters

Tihomir Asparouhov and Bengt Muthén

July 18, 2012

Abstract

Bayesian methodology is well-suited for estimating latent variable models where subjects are not the only random mode, but also items and contexts. A general cross-classified structural equation model is presented where observations are nested within two independent clustering variables. The model includes continuous and categorical dependent variables as well as continuous latent variable. Random effects, intercepts and slopes, are used to model the clustering effects for both nesting structures. We describe the Bayesian methodology implemented in Mplus version 7 used to estimate such models. Bayesian methodology can also be used to estimate cluster specific structural equation models in two-level data where all measurement and structural coefficients, including factor loadings and regression coefficients between factors can be estimated as cluster level random effects rather than fixed parameters. The maximum-likelihood estimation for such models is generally prohibitive due to the large dimension of numerical integration. We also discuss the effect of priors on the Bayesian estimation. In particular we show how a small variance prior can be used to easily identify more random effects than traditional ML methodology can, which can yield flexible structural models with many cluster specific coefficients. Applications are discussed such as multiple group analysis with large number of groups and measurement non-invariance, cross-cultural research and G-theory.

1 Introduction

In this article we describe some of the new modeling possibilities implemented in Mplus version 7 for multilevel and cross classified data. Using the Bayesian methodology it is now possible to substantially expand the set of structural equation models beyond the set of models that can be estimated with the maximum likelihood estimation or the weighted least squares estimation methodology. Currently these two estimators provide the basis for almost all structural equation model fitting.

Data points are typically clustered in several different ways. Fully analyzing the data and accounting for the various clustering effects requires multilevel modeling or cross classified modeling if the clustering effects are not nested within each other. Cross classified modeling has been of interest in the last two decades as the next step of hierarchical modeling, see for example Fielding and Goldstein (2006) for a recent overview.

Cross-classified data arises in various practical applications. One type of applications is the case when the data has two natural clustering variables that are not nested within each other. For example, students are clustered within schools but are also clustered within neighborhoods. Another example is modeling salaries. Individuals are clustered within the same geographical regions and within occupational clusters. Another example is longitudinal students performance scores which are clustered within students and within teachers, see Luo and Kwok (2012). Such a model can be used to simultaneously estimate the students innate abilities and the teachers abilities as well and can be combined with a student level growth model. Another example arises in survey sampling where observations are nested within neighborhoods and interviewers.

A different type of cross-classified applications arises in the analysis of multiple random mode data as discussed in Gonzalez et al. (2008), where observations are nested within persons and cross nested within treatments or situations. In multiple random mode data, there are two or more dimensions that vary randomly. For example, one random dimension can be persons, while the second random dimension can be situational contexts. Persons are random samples from a population. Contexts are also random samples from a target population of contexts. Observations are collected for various persons in various contexts. Experimental designs data can also be treated as multiple random mode data when various experiments are conducted for various subjects. The experiments are randomly selected from a large pop-

ulation of possible experiments and subjects are randomly selected from a population as well.

Finally, a third type of cross-classified applications arises in Generalizability theory, see Cronbach et al. (1963), Marcoulides (1999), Brennan (2001) and De Boeck (2008) where test questions are modeled as random samples from a population of test items. In this case observations are nested within persons and also cross-nested within items.

It should be noted the importance of accounting for all sources of clustering when analyzing clustered data. If one of the clustering effects is ignored, the model is essentially underspecified and fails to discover the true explanatory effect stemming from the additional clustering. This misspecification can also lead to underestimating or overestimating of the standard errors, see Luo and Kwok (2009). One alternative approach to cross-classified modeling is a fixed effects modeling where dummy indicator variables are created for each cluster. If the number of clusters however is more than 10, using fixed effects modeling can lead to too many parameters in the model and that in turn can decrease the accuracy of the estimates and their standard errors.

While multilevel models are easy to estimate through maximum-likelihood via the EM algorithm, see Raudenbush and Bryk (2002) and Goldstein (2011), the cross classified models are not. Rasbash and Goldstein (1994) found a way to obtain the maximum-likelihood estimates for a cross classified model by respecifying the model as a multilevel model. However that method can not be used in general settings when the number of cluster units at both classification levels is large. The MCMC estimation method proposed in Browne et al. (2001) is a general estimation method for fitting cross classified models with no restriction on the data structure. The method can be applied for normally distributed variables as well as categorical variables.

Until recently cross classified modeling discussed in the literature has focused mostly on univariate modeling. The MCMC methodology has now been extended to the multivariate case in Browne et al. (2007) and to the structural equation modeling framework in Gonzalez et al. (2008). In this article, we describe a generalization of the model in Gonzalez et al. (2008) which includes structural equations on all there levels: individual level and the two clustering levels, unbalanced designs in the data, and random slopes for regression coefficients on the individual level. This general cross classified model can be fitted with Mplus version 7 using the MCMC estimation method. The general cross classified framework includes categorical and

continuous dependent variables and continuous latent variables on all three levels.

Another structural equation modeling extension that is made possible by the Bayesian methodology in two-level settings is to allow the measurement loadings to vary across clusters. Traditionally the mean of an observed dependent variable in two-level models varies across clusters and this feature can easily be handled by ML estimation if the variable is continuous or by weighted least squares estimation if the variable is categorical, see Asparouhov and Muthén (2007). However, if the loading varies across clusters the log-likelihood does not have an explicit form and cannot be computed without numerical integration even when the dependent variable is normally distributed. Numerical integration via quadrature would typically not be possible when there are several cluster specific loadings in the model. In such a case the log-likelihood would require the evaluation of a multidimensional integral via numerical quadrature. In practical settings this is not feasible when the dimension is greater than 3 or 4. Thus no more than 3 or 4 random loadings and intercepts can be included in a ML estimated model. In De Jong et al. (2007) and De Jong and Steenkamp (2009) the MCMC methodology has been used to estimate random loading models. The random loadings technique has a number of practical applications. For example, in multilevel structural equation models allowing the random loadings to vary across clusters yields a more flexible model. Another practical application is in the case of multiple group SEM when there are many groups and many minor but significant loadings differences between the groups, see De Jong et al. (2007) and Davidov et al. (2012) for modeling cross-cultural measurement noninvariance. In that case the random loadings approach can be used to obtain a better fitting and a more parsimonious model that allows for loadings non-invariance between the groups and without increasing the number of parameters. In addition, the random loadings model would avoid the step-wise modeling strategy that relaxes loading equalities one at a time. Finally the random loadings feature can be used to allow not only the mean of a variable to vary across clusters but also the variance of a variable to vary across clusters. Such a modeling capability is invaluable for large data sets with many clusters and many observations within clusters where the within level sample variances are significantly different across clusters. The standard multilevel models all assume equal variance across clusters.

2 The general cross classified structural equation model

Suppose that the observed data is clustered within J level 2 units. Suppose also that there is a different type of clustering for this data, we will call this level 3 clustering, with K level 3 units. The data is cross classified when the level 2 clusters are not nested within the level 3 clusters and vice versa, i.e., there is no order implied for the level 2 and the level 3 units.

Suppose that we have a multivariate response vector Y of length P. Denote by Y_{pijk} the $p-$th observed variable for person i belonging to level 2 cluster j and level 3 cluster k. For each level 2 cluster j and level 3 cluster k the number of observations that belong to level 2 cluster j and level 3 cluster k can be any number including 0. In some special experimental designs the number of such observations is exactly 1 and thus the index i in Y_{pijk} is dropped and the data can be presented as Y_{pjk}, see for example Gonzalez et al. (2008). For the purpose of defining the general model however it is not important how many observations are available in the cluster intersection cell (j,k). Raudenbush (1993) shows that the cross classified modeling can be conducted even for sparse situations in which most of the cells are empty. We will assume the general situation where there are any number of observations in that cell. The fact that the two sets of clustering are not nested in each other means that there is at least one level 2 cluster with 2 non-empty cells and at least one level 3 cluster with 2 non-empty cells. Most cross-classified models need at least two non-empty cells for most level 2 and level 3 clusters or at least $J + K$ non-empty cells altogether.

The basic equation for cross classified structural modeling is analogous to the basic multilevel structural equation modeling, see Muthén (1994), where a variable is decomposed as a within cluster variation and between cluster variation. Since there are two level of clustering in cross-classified models the variable is essentially split into three parts. Denote by Y_{ijk} the vector of P observed variables. The basic equation is defined as

$$Y_{ijk} = Y_{1,ijk} + Y_{2,j} + Y_{3,k} \tag{1}$$

where $Y_{2,j}$ is the random effect contribution of the j-th level 2 cluster, $Y_{3,k}$ is the random effect contribution of the k-th level 3 cluster and $Y_{1,ijk}$ is the individual level variation. Alternative interpretation for $Y_{2,j}$ and $Y_{3,k}$ is that they are the random intercepts in the linear model. In this notation we use

a comma to separate the level the variable is defined on and the indices after the comma refer to which individual or cluster the observation is from. Thus the variable $Y_{1,ijk}$ is defined on level 1, $Y_{2,j}$ is defined on level 2, and $Y_{3,k}$ is defined on level 3.

In the above model not all variables need to have all variation components. For example if a variable is defined on level 2 or level 3 then it will have just that corresponding component. Some variables may have cluster contribution from one of the clustering levels but not from the other, see Goldstein (2011) section 12.6. There are 6 possible types of variables in cross classified modeling. One type is having all 3 variations. There are three types having just one variation and there are two types having two variation levels. In the equations below we assume that all three variations are present and if some of the variation is not present it can simply be replaced by 0.

At this point we can define separate structural equations for the 3 sets of variables at the 3 different levels

$$Y_{1,ijk} = \nu_1 + \Lambda_1 \eta_{1,ijk} + \varepsilon_{1,ijk} \tag{2}$$

$$\eta_{1,ijk} = \alpha_1 + B_1 \eta_{1,ijk} + \Gamma_1 x_{1,ijk} + \xi_{1,ijk} \tag{3}$$

$$Y_{2,j} = \Lambda_2 \eta_{2,j} + \varepsilon_{2,j} \tag{4}$$

$$\eta_{2,j} = \alpha_2 + B_2 \eta_{2,j} + \Gamma_2 x_{2,j} + \xi_{2,j} \tag{5}$$

$$Y_{3,k} = \Lambda_3 \eta_{3,k} + \varepsilon_{3,k} \tag{6}$$

$$\eta_{3,k} = \alpha_3 + B_3 \eta_{3,k} + \Gamma_3 x_{3,k} + \xi_{3,k} \tag{7}$$

The variables $x_{1,ijk}$, $x_{2,j}$ and $x_{3,k}$ are the vectors of predictor at the three different levels. The variables $\eta_{1,ijk}$, $\eta_{2,j}$ and $\eta_{3,k}$ are the vectors of the latent variables on the three different levels. The residual variables $\varepsilon_{1,ijk}, \xi_{1,ijk}, \varepsilon_{2,j}, \xi_{2,j}, \varepsilon_{3,k}, \xi_{3,k}$ are zero mean normally distributed residuals with variance covariance matrices $\Theta_1, \Psi_1, \Theta_2, \Psi_2, \Theta_3, \Psi_3$ respectively. In the above equations the intercept parameter vector ν_1 can be present just in one of the three levels. The parameters $\alpha_i, \Lambda_i, B_i, \Gamma_i$ are model parameters to be estimated with one exception. Some or all of the parameters in Γ_1 can be latent variables among the latent variables defined at level 2 and level 3: $\eta_{2,j}$ and $\eta_{3,k}$. These latent variable are the random effect coefficients in the model.

The above model can easily be extended to categorical variable using the probit link function. To use the probit link function for each categorical variable Y_{pijk} we define the underlying variable Y^*_{pijk} such that

$$Y_{pijk} = l \Leftrightarrow \tau_{l-1,p} < Y^*_{pijk} < \tau_{lp} \tag{8}$$

where τ_{lp} are the threshold parameters. In the general model (23-7) we can now substitute the variable Y_{pijk} with Y^*_{pijk}. For identification purposes the variance of $\varepsilon_{1,ijk}$ is fixed to 1.

The estimation of the above model is based on the MCMC algorithm with the Gibbs sampler and is only marginally different from the estimation of the two-level structural equation model described in Asparouhov and Muthén (2010a) and Asparouhov and Muthén (2010b). The two-level model estimation is based on first sampling the between component for each variable from its posterior distribution. Then the Gibbs sampler for two-group structural equation models is used to sample the rest of the components. The two groups are the within and the between levels. Similarly here we sample the two between level components in separate Gibbs sampling steps

$$[Y_{2,j}|*, Y_{ijk}, Y_{3,k}]$$

and

$$[Y_{3,k}|*, Y_{ijk}, Y_{2,j}].$$

Both of these posterior distributions are obtained the same as the posterior for the two-level between components since conditional on $Y_{3,k}$ the model for Y_{ijk} is essentially a two-level model with the between component being $Y_{2,j}$. Similarly the conditional posterior distribution of $Y_{3,k}$ given $Y_{2,j}$ is the same as the posterior for the two-level component in a two-level model. After the above two steps, the Gibbs sampler continues sampling the remaining components as if the model is a 3 group structural equation model where $Y_{1,ijk}$, $Y_{2,j}$ and $Y_{3,k}$ are the observed variables in the 3 groups.

2.1 Cross classified factor analysis model

In this section we present a simulation study for a cross-classified factor analysis model with 5 observed variables. We generate data with M clusters at level 2 and M clusters at level 3. In every cell intersection there is only one observation, i.e., there is exactly one observation belonging to level 2 cluster

j and level 3 cluster k. Thus the sample size is M^2. The factor analysis model has one factor at each level and can be described by the following equation

$$y_{pijk} = \mu_p + \lambda_{1,p} f_{1,ijk} + \lambda_{2,p} f_{2,j} + \lambda_{3,p} f_{3,k} + \varepsilon_{2,pj} + \varepsilon_{3,pk} + \varepsilon_{1,pijk}.$$

In this model we fix all factor variances to 1 and all loadings are estimated. The parameter values we use for generating the data are as follows $\lambda_{1,p} = 1.5$, $\theta_{1,p} = 1.2$, $\lambda_{2,p} = 0.8$, $\theta_{2,p} = 0.5$, $\lambda_{3,p} = 1$, $\theta_{3,p} = 1.5$, $\mu_p = 2.2$. In this simulation study we vary the number of cluster units M to evaluate the performance of the Bayesian estimator. We generate 100 samples for each value of M.

In Table 1 we present the absolute bias and the coverage for some of the model parameters. The results show that for $M = 10$ there are some substantial biases for the between level parameters. For $M = 20$ there are some moderate biases and for $M = 30$ or more the biases are small. In all cases the coverage is near the nominal value of 95%. These results indicate that when the number of clustering units is small we can expect biases in the estimates on that clustering level. In a different simulation study, not reported here we increase the number of units in each cross classified cell from 1 to 10. This however that did not affect the biases on the between levels but reduced the biases on the within level.

We can conclude from this simulation study that in asymptotic settings when the number of cluster level units is large at both clustering levels the Bayes estimator yields consistent results. On the other hand the Bayes estimator is much more flexible than the ML estimator in small sample size estimation due to the fact that we can add informative or slightly informative priors to the estimation. When the sample size is small these priors will affect the results. To illustrate this point we show how the biases in the case of $M = 10$ can be substantially reduced by providing weakly informative priors. For all parameters on the two between levels we set weakly informative priors that approximately match the mean and the variance of the posterior from the Bayes estimation based uninformative priors. This approach can be viewed as a two-stage estimation where the first stage is used only to get a more narrow range of the estimates. In our example, using conjugate priors, we set the prior of the loadings on the second level to $N(1.1, 1.5)$ and the loadings on the third level to $N(1.4, 3.5)$. The priors for the variance parameters on the second level are set to $IG(2.3, 1)$ and on the third level to

Table 1: Absolute bias and coverage for cross-classified factor analysis model

Parameter	M=10	M=20	M=30	M=50	M=100
$\lambda_{1,1}$	0.07(0.92)	0.03(0.89)	0.01(0.95)	0.00(0.97)	0.00(0.91)
$\theta_{1,1}$	0.05(0.96)	0.00(0.97)	0.00(0.95)	0.00(0.99)	0.00(0.94)
$\lambda_{2,p}$	0.21(0.97)	0.11(0.94)	0.10(0.93)	0.06(0.94)	0.00(0.92)
$\theta_{2,p}$	0.24(0.99)	0.10(0.95)	0.04(0.92)	0.05(0.94)	0.02(0.96)
$\lambda_{3,p}$	0.45(0.99)	0.10(0.97)	0.03(0.99)	0.01(0.92)	0.03(0.97)
$\theta_{3,p}$	0.75(1.00)	0.25(0.98)	0.15(0.97)	0.12(0.98)	0.05(0.92)
μ_p	0.01(0.99)	0.04(0.98)	0.01(0.97)	0.05(0.99)	0.00(0.97)

$IG(2.3, 3)$. Table 2 shows the effect of adding these priors. The biases are reduced substantially. This two-stage approach has been effective in other situations, see Asparouhov and Muthén (2010b). The priors that are constructed from the first stage estimation are considered weakly informative prior.

In the above simulation the number of level 2 clusters was set to be equal to the number of level 3 clusters. In practical applications however that will not be the case. For example, in Gonzalez et al. (2008), the number of level 2 clusters was 679 while the number of level 3 clusters was 11. The large number of level 2 clusters can ensure that the level 2 parameter estimates are unbiased, however the low number of level 3 units suggest that the level 3 parameter estimates may have biases similar to those obtained in our simulation study for the case $M = 10$. This will be true in general. The quality of the estimates on a particular level is driven primarily by the number of cluster units on that level. The second stage estimation based on weakly informative priors should be used only for the levels with small number of cluster units.

Table 2: Absolute bias and coverage for cross-classified factor analysis model and small number of clusters

Parameter	M=10 with uninformative priors	M=10 with first stage priors
$\lambda_{1,1}$	0.07(0.92)	0.07(0.92)
$\theta_{1,1}$	0.05(0.96)	0.05(0.97)
$\lambda_{2,p}$	0.21(0.97)	0.16(0.98)
$\theta_{2,p}$	0.24(0.99)	0.05(1.00)
$\lambda_{3,p}$	0.45(0.99)	0.28(0.94)
$\theta_{3,p}$	0.75(1.00)	0.12(1.00)
μ_p	0.01(0.99)	0.00(0.97)

2.2 Cross classified path analysis with binary variables

In this section we conduct a simulation study based on the path analysis model described in Gonzalez et al. (2008). The model has 4 binary variables: Frustration, Irritation, Anger and Antagonistic tendency. In this application the observations are nested within person and within situations. As in the previous section we will vary the number of clusters at the two clustering levels to study the quality of the estimates when the number of cluster units is small. For each pair of level 2 and level 3 clusters there is just 1 observation in common. The model is described as follows

$$y^*_{pjk} = y_{2,pj} + y_{3,pk} + \varepsilon_{1,pjk}$$

where as usual the variances of $\varepsilon_{1,pjk}$ is fixed to 1 for identification purposes. The structural model is the following path analysis model that uses the same structural coefficients on both between levels

$$y_{2,1j} = \beta_1 y_{2,3j} + \beta_2 y_{2,4j} + \varepsilon_{2,1j}$$

$$y_{2,2j} = \beta_3 y_{2,3j} + \beta_4 y_{2,4j} + \varepsilon_{2,2j}$$

$$y_{3,1k} = \beta_1 y_{3,3k} + \beta_2 y_{3,4k} + \varepsilon_{3,1k}$$

$$y_{3,2k} = \beta_3 y_{3,3k} + \beta_4 y_{3,4k} + \varepsilon_{3,2k}.$$

Covariance parameters are estimated in this model between the following pairs of variables $Cov(\varepsilon_{2,1j}, \varepsilon_{2,2j}) = \psi_{2,12}$, $Cov(\varepsilon_{3,1k}, \varepsilon_{3,2k}) = \psi_{3,12}$, $Cov(y_{2,3j}, y_{2,4j}) =$

Table 3: Absolute bias and coverage for cross classified path analysis with binary variables

Parameter	M=10	M=20	M=30	M=50	M=100
β_1	0.13(0.92)	0.05(0.89)	0.00(0.97)	0.01(0.92)	0.01(0.94)
$\psi_{2,11}$	0.11(1.00)	0.06(0.96)	0.01(0.98)	0.00(0.89)	0.02(0.95)
$\psi_{2,12}$	0.15(0.97)	0.06(0.92)	0.05(0.97)	0.03(0.87)	0.01(0.96)
τ_1	0.12(0.93)	0.01(0.93)	0.00(0.90)	0.03(0.86)	0.00(0.91)

$\psi_{2,34}$, $Cov(y_{3,3k}, y_{3,4k}) = \psi_{3,34}$. The true values used for the data generation are as follows: $\psi_{2,11} = \psi_{2,33} = 0.5$, $\psi_{2,22} = \psi_{2,44} = 0.8$, $\psi_{2,12} = \psi_{2,34} = 0.4$, $\psi_{3,11} = \psi_{3,33} = 1.5$, $\psi_{3,22} = \psi_{3,44} = 0.9$, $\psi_{3,12} = \psi_{3,34} = 0.7$, $\beta_1 = 0.3$, $\beta_2 = -0.3$, $\beta_3 = 0.5$, $\beta_4 = -0.5$, $\tau_1 = \tau_3 = 0.2$, $\tau_2 = \tau_4 = -0.5$.

Table 3 contains the results of the simulation study for some of the model parameters. Again we see that when $M = 10$ the estimates are somewhat biased and as the number of clusters increases the biases disappear. For $M = 10$ a two-stage estimation can be used to reduce the biases.

It is interesting to note that the Gonzalez et al. (2008) model does not contain a level 1 model, while the general model presented in this article, see (23-7), includes such a model. To check the quality of the estimation for models that include a level 1 model we conduct a simulation study that includes an unrestricted correlation matrix on level 1 for the variables $\varepsilon_{1,1jk}, ..., \varepsilon_{1,4jk}$. The results of this simulation study are very similar to those in Table 3 and thus are not presented. It should be noted that the level 1 parameters are the easiest to estimate. These parameters typically have the smallest standard errors and there is more information in the data about these parameters than there is for the between level parameters even when the intersection cells contain just one observation or the cross-sectional data is sparse. The level 1 parameters are also much less affected by having a small number of between level units M. Thus the model (23-7) provides a valuable additional level of flexibility. Omitting the level 1 model could result in model misspecification and biased parameter estimates.

3 Two level structural equation models with random loadings

In this section we show how the Bayesian methodology can be used to estimate a generalized two-level structural equation model which has the flexibility of cluster specific loadings. Let Y_{ij} be the vector of observed variables. The general two-level structural equation model is defined by the following equations

$$Y_{ij} = Y_{1,ij} + Y_{2,j} \tag{9}$$

$$Y_{1,ij} = \Lambda_1 \eta_{1,ij} + \varepsilon_{1,ij} \tag{10}$$

$$\eta_{1,ij} = B_1 \eta_{1,ij} + \Gamma_1 x_{1,ij} + \xi_{1,ij} \tag{11}$$

$$Y_{2,j} = \nu_2 + \Lambda_2 \eta_{2,j} + \varepsilon_{2,j} \tag{12}$$

$$\eta_{2,j} = \alpha_2 + B_2 \eta_{2,j} + \Gamma_2 x_{2,j} + \xi_{2,j} \tag{13}$$

The variables $x_{1,ij}$ and $x_{2,j}$ are the vectors of predictor variables observed at the two levels. The variables $\eta_{1,ij}$ and $\eta_{2,j}$ are the vectors of latent variables. The residual variables $\varepsilon_{1,ij}, \xi_{1,ij}, \varepsilon_{2,j}, \xi_{2,j}$ are zero mean normally distributed residuals with variance covariance matrices $\Theta_1, \Psi_1, \Theta_2, \Psi_2$ respectively. The parameters $\alpha_i, \Lambda_i, B_i, \Gamma_i$ are the model parameters to be estimated. The above model can easily be estimated in Mplus using the maximum-likelihood estimator or with the Bayes estimator. The model can also be extended to categorical variables as in (8). In that case the model can be estimated with the weighted least squares estimator or the Bayes estimator. If the variables are all normally distributed the regression parameters Γ_1 can be estimated as between level random effects. The ML estimator can easily estimate such a model as well. However the remaining within level parameters Λ_1, B_1, Θ_1 and Ψ_1 cannot been allowed to vary across clusters within the standard ML estimation framework. Such an estimation is possible in principle, but it requires numerical integration. In many cases the numerical integration cannot be performed accurately or it is too heavy computationally. The Bayesian methodology in contrast can very easily be extended to allow Λ_1, B_1, Θ_1 and Ψ_1 to vary across clusters. For now we will

focus on Λ_1 and B_1 but later we will discuss how Θ_1 and Ψ_1 can also vary across clusters.

Consider the following generalization of the two-level structural equation model (10-13). In this model not only Γ_1 but also Λ_1 and B_1 are allowed to vary across clusters, i.e., these parameters are set equal to the random effects in the vector $\eta_{2,j}$. The Bayesian estimation of this model is again a marginal extension of the standard two-level SEM estimation as described in Asparouhov and Muthén (2010a) and (2010b). Conditional on the within level latent variable the Gibbs sampler for the between level random effects is the same as if the within level latent variable is observed, i.e., once the within level latent variable is sampled the conditional posteriors are as in the regular SEM model with random effects on observed variables.

This generalization can also be extended to the exploratory factor analysis (EFA) model and the exploratory structural equation model (ESEM), see Asparouhov and Muthén (2009). The estimation of the EFA model is based on first estimating the unrotated model, which essentially is a confirmatory factor analysis model with minimal number of restrictions. Using the cluster specific loading model above, the unrotated solution can be estimated as a cluster specific random effect model for all unrotated loadings. After convergence of the MCMC sequence the posterior distribution of the cluster specific unrotated solution is obtained. This unrotated posterior distribution is then rotated one MCMC iteration at a time to obtain the cluster specific posterior distribution of the rotated solution. The estimation is similar to the Bayesian estimation of the basic EFA model, see Asparouhov and Muthén (2012). This estimation is prone to sign and permutation switching for the unrotated and rotated model and additional constraints are needed for proper estimation. Such constraints can also be enforced on cluster specific level.

3.1 Comparing Bayesian and ML-Montecarlo estimations on a random loadings factor analysis model

In this section we compare the Bayesian estimation and the ML estimation with Montecarlo integration on a simple factor analysis model with random loadings. The model has 5 continuous indicators and 1 factor on the within level and is described by the following equation

$$Y_{pij} = \mu_{pj} + \lambda_{pj}\eta_{ij} + \varepsilon_{ijk} \tag{14}$$

Table 4: Absolute bias and coverage for factor analysis model with random loadings - comparing Bayes v.s. ML-Montecarlo

parameter	Bayes	Monte 500	Monte 5000
$\theta_{1,1}$	0.00(0.97)	0.65(0.00)	0.42(0.01)
$\nu_{2,1}$	0.01(0.95)	0.01(0.78)	0.00(0.80)
$\alpha_{2,1}$	0.01(0.96)	0.08(0.50)	0.04(0.60)
$\theta_{2,1}$	0.02(0.89)	0.23(0.31)	0.15(0.50)
$\psi_{2,1}$	0.02(0.91)	0.10(0.23)	0.10(0.21)

for $p = 1, ...5$. The between level random effects α_{pj} and λ_{pj} are assumed independent among each other and have means and variance $\nu_{2,p}$ and $\theta_{2,p}$, and $\alpha_{2,p}$ and $\psi_{2,p}$ respectively. The variance of η_{ij} is fixed to 1 for identification purposes and the variance of ε_{ijk} is $\theta_{1,p}$. We generate 100 samples with 100 clusters of size 10 and analyze them with the Bayes estimator as well and with MLE based on Montecarlo integration with 500 points of numerical integration and also 5000 points of numerical integration. All three methods are implemented in Mplus. The Montecarlo integration essentially is applied to 11 dimensional integration in this example because there are 10 between level random effects and one within level factor in the model. To generate the data we use the following parameter values: $\nu_{2,p} = 0$, $\theta_{2,p} = 0.6$, $\alpha_{2,p} = 1$, $\psi_{2,p} = 0.2$, $\theta_{1,p} = 1$. Table 4 contains the bias and coverage for the 3 estimation methods for some of the model parameters. The results clearly illustrate the advantage of the Bayesian estimation method. The bias for the Bayesian method is nearly 0 for all parameters and coverage is near the nominal level of 95%. On the other hand the ML method shows bias in the point estimates and the coverage of the parameter estimates is very low. Increasing the number of integration points improves the estimates but not sufficiently in comparison to the Bayesian estimation.

3.2 The effect of treating random loadings as fixed parameters

In this section we study the consequences of ignoring the randomness of the loadings, i.e., we determine the result of incorrectly assuming that the loadings are the same across clusters. Using the example from the previ-

Table 5: Absolute bias and coverage for factor analysis model with random loadings - comparing random loadings v.s. fixed loadings models

parameter	Bayes	ML with fixed loadings
$\theta_{1,1}$	0.00(0.97)	0.20(0.23)
$\nu_{2,1}$	0.01(0.95)	0.14(0.66)
$\alpha_{2,1}$	0.01(0.96)	0.00(0.80)
$\theta_{2,1}$	0.02(0.89)	0.00(0.93)
$\psi_{2,1}$	0.02(0.91)	-

ous section we compare the Bayesian estimates which take into account the variability of the loadings across cluster and the ML estimates for the model that assumes equal loadings across clusters, i.e., the ML estimator is applied to the the model where the parameter $\psi_{2,p}$ is fixed to 0. Table 5 contains the comparison between the two methods. The results indicate as expected that some of the ML estimates are biased. It is also interesting to note that the loadings parameters $\alpha_{2,1}$ are not biased but their standard errors are underestimated and the coverage for the loadings drops down to 80%.

Next we conduct a simulation study to determine the effect of ignoring the variability in the loadings in models with categorical data. In this case we compare the Bayesian estimates accounting for the variability in the loadings with the weighted least squares estimates assuming fixed loadings. We use again the model described in (14) where now the dependent variable is really Y^* instead of Y. We generate 5 category variables using the threshold values ± 0.5 and ± 2. Again we generate 100 samples with 100 clusters of size 10. We use the following parameter values to generate the data $\nu_{2,p} = 0$, $\theta_{2,p} = 0.6$, $\alpha_{2,p} = 1$, $\psi_{2,p} = 0.3$, $\theta_{1,p} = 1$. The results are presented in Table 6. The Bayes estimates have a small bias and a good coverage while the weighted least squares estimates have a larger bias and a poor coverage for all parameters including the loading parameters.

Table 6: Absolute bias and coverage for factor analysis model with categorical data and random loadings - comparing random loadings v.s. fixed loadings models

parameter	Bayes	WLSMV with fixed loadings
$\tau_{1,1}$	0.05(0.96)	0.17(0.63)
$\alpha_{2,1}$	0.03(0.92)	0.13(0.39)
$\theta_{2,1}$	0.05(0.91)	0.11(0.70)
$\psi_{2,1}$	0.05(0.89)	-

3.3 Factor model for the random loadings

In the previous two sections we considered an example where the between level variations are independent among each other. This however is not a realistic assumption. To resolve this problem in model (14) we can assume that all between level random effects are correlated. This approach will contribute 45 new parameters to the model. A more parsimonious approach is to model the correlations between the random effects through a factor analysis model on the between level. This will reduce the number of additional parameters in the model down to 20. This model can be written as follows. In addition to equation (14) the between level factor model is given by

$$\mu_{pj} = \mu_p + b_{2,p}\eta_{2,j} + \varepsilon_{2,p} \tag{15}$$

$$\lambda_{pj} = \lambda_p + b_{2,P+p}\eta_{2,j} + \varepsilon_{2,P+p} \tag{16}$$

where the variance of $\eta_{2,j}$ is fixed at 1. We generate 100 data sets according to this model with 5 indicator variables and 100 clusters of size 10. We use the following parameter values to generate the data $\mu_p = 0$, $\psi_{2,p} = 0.6$, $\psi_{2,P+p} = 0.3$, $\theta_{1,p} = 1$, $\lambda_p = 1$, $b_{2,p} = 0.6$, $b_{2,P+p} = 0.4$. We then analyze the data with the Bayesian estimator and the model given by (14) together with (15) and (16), i.e., with the correct model. We also analyze the data with the model where μ_{pj} and λ_{pj} are assumed to be independent, i.e., using only model (14). Table 7 contains the bias and coverage for a selection of the parameters and it is clear from these results that the Bayesian estimation works quite well when the correlations between the level 2 random effects are taken into account. When the random effects are estimated as independent some biases do occur and the coverage drops. Thus we conclude that it is

Table 7: Absolute bias and coverage for factor analysis model with correlated random loadings

parameter	Random effect factor model	Independent random effects
$\theta_{1,1}$	0.01(0.97)	0.01(0.97)
$b_{2,1}$	0.04(0.95)	-
$b_{2,6}$	0.04(0.94)	-
μ_1	0.00(0.98)	0.07(0.84)
λ_1	0.01(0.97)	0.05(0.90)
$\psi_{2,1}$	0.03(0.92)	0.14(0.83)
$\psi_{2,6}$	0.03(0.92)	0.04(0.92)

important to estimate the correlations between the level 2 random effects even though that may require less parsimonious model and more computational time.

3.4 Loadings with large variance and negative loadings

In De Jong et al. (2007) and De Jong and Steenkamp (2009) all loadings are restricted to be positive. Negative loadings however do occur in practical applications and thus it is important to evaluate the performance of the estimation method in the presence of negative loadings. In this section we conduct two simulation studies where some of the loadings are negative for some clusters due to a negative mean or a large variance for a random loading. We modify the simulation study described in Section 3.1. In the first modification which we denote by M1 all random loading variances are increased to 1. Thus the variability of the random loadings is increased and the range of the loadings is between -1 and 3. In the second modification, denoted by M2, the mean of the first loading is set to -1 and the variance is set to 1. Table 8 contains the results of these two simulation studies. It is clear from the results that the MCMC estimation method performs well. However, 7% of the replications did not converge in simulation M1 and 2% of the replications did not converge in M2 using 50000 MCMC iterations and the default PSR convergence criteria in Mplus. Such convergence problems most likely can be resolved by using more MCMC iterations.

Factor analysis models have a sign non-identification. Reversing the sign

Table 8: Absolute bias and coverage for factor analysis model with random loadings - large variance and negative loadings

parameter	M1	M2
$\theta_{1,1}$	0.00(0.95)	0.01(0.93)
$\nu_{2,1}$	0.01(0.97)	0.01(0.92)
$\alpha_{2,1}$	0.01(0.94)	0.03(0.95)
$\theta_{2,1}$	0.01(0.93)	0.03(0.94)
$\psi_{2,1}$	0.06(0.90)	0.03(0.92)

of all factor loadings produces an equivalent model. In Bayesian estimation of factor analysis models the posterior distribution is bimodal due to this sign non-identification. In the MCMC estimation it is important to sample only from one of the two symmetric parts of the posterior distribution of the loadings otherwise the median and the mean of the posterior will be zero and the results will not be meaningful. One standard way of resolving this problem is to constraint the generated loadings so that the sum of all loadings for a given factor is always positive. This essentially will split the posterior distribution in two symmetric parts. Alternatively in each MCMC step if the sum of the loadings is negative the sign of all loadings can be reversed and the MCMC sequence can continue sampling only from one of the two symmetric modes. When we are estimating a random loadings model the situation is similar. Here the sign of the mean of the random loadings can be reversed. To ensure again that only one mode of the posterior distribution is sampled one can introduce a parameter constraint such that the sum of the means of all random loadings for each factor is positive.

3.5 Small number of clusters

In many practical applications the number of groups or clusters is small, the size of the clusters is large, and the number of variables may be large as well. When we estimate a simple one factor analysis model with cluster specific mean and cluster specific loading the model has $2P$ between level random effects. To estimate a variance covariance matrix of size $2P$ generally we need at least $2P$ observations on the between level, i.e., we need at least $2P$ clusters. It is important to know how the random loading model performs

when the number of clusters is small, i.e., smaller than $2P$ or slightly larger than $2P$.

One way to deal with small number of clusters is to add an informative prior for the variance covariance matrix on the between level. In many situations however such prior is not available. In the simulations below we will use the Mplus default uninformative priors for the variance parameters. For the variance parameter the Mplus default prior is the inverse gamma prior $IG(-1, 0)$.

Alternative way to deal with the small number of clusters problem is to estimate the random effects variance covariance matrix as a restricted variance covariance. Two models will be considered here. The first model is the uncorrelated model where all random effects on the between level are assumed independent. The second model is the factor analysis model where a one factor model is estimated for the between level random effects. We conduct a simulation study with both binary and normally distributed variables. In all cases we generate the data and analyze it with the correct model. Simulation study is conducted with different number of observed variables P and different number of clusters M. The number of variable P is 5, 10, 20, 30 or 50. The number of clusters M is 10, 20, or 30. The size of the clusters is 100 is all simulations. The parameter values used for this simulation are as follows. The factor variances are fixed to 1, the residual variances are set to 1, the random intercepts have mean 0 and variance 0.6, the random loadings have mean 1 and variance 0.3. In the binary case the threshold values are set to 0. For the 1-factor analysis model on the between level the loadings for the random intercepts are set to 0.6 and for the random loadings to 0.3.

Another way to deal with the small number of clusters problem is to use the BSEM methodology developed in Muthén and Asparouhov (2012) which allows us to use unidentified structural models and tiny informative priors to resolve the unidentifications. The BSEM methodology can be used to analyze data with more flexible models that reflect more accurately the real data applications. For example, in a factor analysis model the residual correlations between the measurement variables are typically assumed to be zero. This assumption however is not realistic in many applications. The BSEM methodology allows us to estimate the factor analysis model under the assumption that the residual correlations are approximately zero by specifying small priors for the correlations that are peaked at zero but allow for some wiggle room for these parameters if there is information in the data

that suggest the residual correlations are not exactly zero. This modeling approach reflects more accurately the believe that the correlations should be small if not zero. To resolve the small number of clusters problem however we need to estimate a model for the between level random effect that is somewhat different than those discussed in Muthén and Asparouhov (2012). This is because our prior assumption for the between level random effect is not that they are independent as in Muthén and Asparouhov (2012) but that they are highly correlated. In most practical applications the cluster level random effects are correlated. To incorporate such a prior believe in the model we estimate a 1 factor analysis model on the cluster level with the factor variance and all loadings parameters estimated as a free parameters. The loading parameters however have an informative prior distribution of $N(1, \sigma)$ where the variance parameter σ is chosen to be a small value. If σ is chosen to be 0 then essentially all loading parameters are fixed to 1. To allow however for a more flexible correlation matrix on the between level where not all covariances are the same between the random effects the prior variance σ is chosen to be a positive value. Typically the model is estimated first with a small value such as $\sigma = 0.01$ and the value is increased until the loading estimates stop changing or until the model stops being identified. If σ is chosen to be a large value then the prior is essentially uninformative and the model will be unidentified because both the factor variance and the loading parameters are estimated. The correlations between the cluster level random effect are driven primarily by the free factor variance parameters and the factor loadings are all expected to be near 1. The variations from 1 are essentially used to more accurately model the correlations between the random effects. In our simulation we choose the value of σ to be 0.5. With this choice the estimated loadings preserve approximately the ration of 2:1 as in the true model and therefore the between level correlations are properly accounted for. This BSEM 1-factor model is essentially a middle ground between the uncorrelated random effect model and the 1-factor analysis model. the model has the advantage over the uncorrelated model that it allows for random effect correlations to be estimated and that would lead to more accurate estimation of the random effects. The 1-factor BSEM model also has an advantage over the 1-factor model since it requires fewer number of clusters to be estimated. In general multilevel modeling the number of random effects that can be estimated by the ML estimator is quite limited. Practical applications with more than 3 or 4 random effects are very few. This is primarily due to the fact that between level random effects variance covariance

matrices become singular and the random effects become perfectly correlated in the estimation. The BSEM 1-factor analysis model generally resolves this problem and can be used to estimate any number of random effects. In pour simulation for example the BSEM 1-factor analysis model successfully estimates a model with 100 random effects, in the case of $P = 50$ even when the number of clusters is 10.

Table 9 shows the convergence rates for the Bayesian estimation method. For the uncorrelated model we obtained 100% convergence in all cases. For the 1-factor analysis model we obtained good convergence rates only when the number of clusters M is bigger than then number of random effects $2P$. The BSEM 1-factor analysis model also converged near 100% of the time in all cases.

When the number of clusters M is 10 the variances of the random effects are generally biased and dependent on the priors. Table 10 shows the bias and coverage for the variance parameter of the first random loading for $P = 5$ and $P = 20$ for the uncorrelated model with normal variables. In practical applications with 10 clusters if an informative prior is not available for these variance parameters the estimation should be conducted in two stages as described in Section 2.1. The first stage should be used to obtain a appropriate range for the prior, i.e., it should be used to construct a weakly informative prior which will then be used in the second stage estimation. The variance of the random intercept in these simulations also show biases when the number of clusters is 10.

Table 9: Convergence rate for random factor analysis model with small number of clusters

Variables	Model	Number of Variables	M=10	M=20	M=30
Normal	Uncorrelated	5	100%	100%	100%
Normal	Uncorrelated	10	100%	100%	100%
Normal	Uncorrelated	20	100%	100%	100%
Normal	Uncorrelated	30	100%	100%	100%
Normal	Uncorrelated	50	100%	100%	100%
Binary	Uncorrelated	5	100%	100%	100%
Binary	Uncorrelated	10	100%	100%	100%
Binary	Uncorrelated	20	100%	100%	100%
Binary	Uncorrelated	30	100%	100%	100%
Binary	Uncorrelated	50	100%	100%	100%
Normal	1-factor	5	0%	96%	98%
Normal	1-factor	10	0%	0%	96%
Normal	1-factor	20	0%	0%	0%
Normal	1-factor	30	0%	0%	0%
Normal	1-factor	50	0%	0%	0%
Normal	BSEM 1-factor	5	100%	96%	98%
Normal	BSEM 1-factor	10	100%	100%	98%
Normal	BSEM 1-factor	20	100%	100%	98%
Normal	BSEM 1-factor	30	100%	100%	98%
Normal	BSEM 1-factor	50	100%	100%	100%

Table 10: Bias and coverage for the variance of the random loading

Number of Variables	M=10	M=20	M=30
5	0.11(0.98)	0.03(0.92)	0.03(0.92)
20	0.10(0.94)	0.04(0.98)	0.02(1.00)

3.6 Cluster specific variance

In this section we demonstrate how random loading can be used to generalize any multilevel model to include cluster specific variance. Consider the simple two-level univariate model

$$Y_{ij} = Y_{2,j} + \varepsilon_{ij}$$

where $Y_{2,j}$ is the between level random intercept with mean parameter μ and variance parameter θ_2. The standard assumption for multilevel models is that ε_{ij} has the same variance θ_1 across clusters. Thus the conditional distribution of $Y_{ij}|j$ is a normal distribution with mean $Y_{2,j}$ and variance θ_1, i.e., only the mean is allowed to vary across clusters in standard multilevel models. When the clusters are large however and there are many clusters in the data the constant variance assumption might not be realistic. The random loading model can easily facilitate non-constant variance using the following equation

$$\varepsilon_{ij} = \lambda_j \eta_{ij}$$

where the variance of the factor η_{ij} is 1 and λ_j is a cluster level random effect with mean α and variance ψ. We conduct a simulation study to evaluate the performance of this model. We generate 100 data sets with 100 clusters of size 10 using the following parameter values $\mu = 0, \theta_2 = 1.6, \alpha = 2, \psi = 0.5$. We analyze the data using the true model that allows cluster specific variance and also the model that assumes constant variance across clusters. The results are presented in Table 9. The only difference in the results between the two methods is in the standard error for θ_1. Incorrectly assuming that the variance is the same across clusters leads to underestimation of the standard error for θ_1 and low coverage. In a more complicated model the bias in the standard error for θ_1 can affect the standards errors for other other parameters as well.

There are two other extensions that should be noted. It is easy to explore any correlation between the cluster specific mean and the cluster specific variance. In the above framework we can simply estimate the correlation between the two cluster random effects $Y_{2,j}$ and λ_j. It should also be noted that the Bayesian estimation can be used to estimate multivariate models with many observed or latent variables having cluster specific variances and covariances. This is another advantage of the Bayesian estimation over the ML estimation which allows for a limited number of two-level random effects when random loadings are present in the model.

Table 11: Absolute bias and coverage - cluster specific variance

parameter	Cluster specific variance	Constant variance
μ	0.01(0.91)	0.02(0.93)
θ_1	0.04(0.90)	0.04(0.73)
θ_2	0.11(0.88)	0.10(0.88)

3.7 Student evaluation of teacher effectiveness example

In this section we describe a practical application of the random loading model. We use the SEEQ data, student evaluation of teacher effectiveness, described in Marsh and Hocevar (1991). The data consists of 35 continuous items. The teacher evaluations are split in 21 subsample based on the qualifications of the teacher and the academic discipline. In Marsh and Hocevar (1991) a 9 factor analysis model is considered. For simplicity however, we consider a 1 factor analysis model, although using several factors does not elevate the complexity of the model particularly when all measurements load on a single factor. It was noted in Marsh and Hocevar (1991) that minor variation exists in the loadings between the 21 groups. The sample size in this application is 24158 and therefore any minor variations in the factor loadings between the groups would be statistically significant. Thus the model that imposes measurement invariance would be rejected. Such sample size and group combinations are not unusual. Davidov et al. (2012) analyzes data from 26 countries and 43779 observations.

The model without measurement invariance would have more than 1500 parameters and thus would not be parsimonious. It is possible to evaluate the measurement non-invariance for each variable and group however that would be a very tedious process given that over 1500 parameters are involved and it is not clear which subset of parameters should be held equal and which should not be held equal. If the measurement non-invariance is ignored the factor score estimates which represents the teacher effectiveness could have a substantial error. One method for dealing with measurement non-invariance using the Bayesian methodology is developed in Muthén and Asparouhov (2012) where all parameters are estimated but approximate equality is enforced between the loadings across groups, i.e., the loadings are estimated

as different parameters but a strong prior specification is used that amounts to holding the loadings approximately equal. This approach treats the loadings as fixed parameters rather than random effects. In this section we will illustrate how the random loading model can be used to easily resolve the measurement invariance problems.

To illustrate the performance of the random loading model we standardize all variables and estimate four different models. The first model M1 is the standard 1 factor analysis model estimated with the maximum-likelihood estimator. The second model M2 is a 1-factor analysis model with uncorrelated random intercepts estimated with the maximum-likelihood estimator. The third model M3 is a 1-factor analysis model with uncorrelated random intercepts and loadings using the Bayes estimator. The fourth model M4 is a 1-factor analysis model with BSEM 1-factor analsyis model for the random intercepts and loadings on the clusters level. The prior for the cluster level loadings is set to $N(1, 0.1)$.

Table 12 contains the factor loadings estimates for the 35 variables and the three models. In Models M3 and M4 we also include the 95% range for the cluster specific factor loadings based on the estimated normal distribution for the random loadings. It can be seen from these results that the random loading range is quite wide and that cluster specific loadings can be substantially different from their fixed ML based estimates. In the 1-factor BSEM model the average R^2 for the 35 random loadings is 0.2, i.e., the correlations between the random loadings are not very high. We conclude that the random loadings model can be used to elegantly accommodate the factor loading variation across groups.

When measurement invariance is not present we can still compare the factor means across groups. This can be done for example by estimating a random intercept model for the individual level factor.

Table 12: Factor loading estimates for SEEQ data

M1	M2	M3	M4
0.76	0.79	0.81(0.67,0.94)	0.79(0.63,0.94)
0.78	0.81	0.82(0.71,0.92)	0.80(0.69,0.91)
0.80	0.80	0.82(0.71,0.93)	0.80(0.69,0.91)
0.73	0.74	0.76(0.67,0.85)	0.74(0.65,0.83)
0.88	0.83	0.85(0.70,1.01)	0.83(0.70,0.96)
0.89	0.83	0.85(0.74,0.96)	0.83(0.74,0.93)
0.84	0.75	0.77(0.64,0.91)	0.75(0.63,0.87)
0.90	0.87	0.89(0.83,0.95)	0.87(0.81,0.93)
0.81	0.85	0.87(0.73,1.01)	0.85(0.73,0.98)
0.78	0.85	0.87(0.71,1.02)	0.85(0.69,1.00)
0.78	0.83	0.84(0.71,0.98)	0.83(0.68,0.97)
0.75	0.71	0.73(0.58,0.88)	0.71(0.54,0.89)
0.80	0.66	0.67(0.48,0.85)	0.65(0.49,0.82)
0.79	0.65	0.66(0.47,0.86)	0.65(0.47,0.83)
0.86	0.80	0.82(0.68,0.95)	0.80(0.68,0.93)
0.84	0.73	0.75(0.59,0.90)	0.73(0.59,0.87)
0.73	0.75	0.77(0.57,0.97)	0.75(0.58,0.92)
0.70	0.75	0.77(0.60,0.93)	0.75(0.60,0.90)
0.76	0.77	0.80(0.66,0.93)	0.77(0.65,0.90)
0.66	0.69	0.72(0.58,0.85)	0.70(0.57,0.83)
0.82	0.76	0.78(0.64,0.92)	0.76(0.62,0.90)
0.84	0.77	0.79(0.64,0.94)	0.77(0.62,0.92)
0.86	0.78	0.80(0.65,0.95)	0.78(0.64,0.92)
0.80	0.73	0.75(0.63,0.87)	0.73(0.62,0.84)
0.70	0.77	0.79(0.66,0.91)	0.77(0.63,0.90)
0.74	0.78	0.80(0.62,0.98)	0.79(0.61,0.96)
0.73	0.78	0.79(0.65,0.93)	0.77(0.63,0.92)
0.62	0.63	0.64(0.44,0.84)	0.62(0.42,0.83)
0.68	0.73	0.74(0.57,0.90)	0.72(0.55,0.89)
0.83	0.89	0.90(0.77,1.04)	0.89(0.73,1.04)
0.92	0.92	0.93(0.84,1.02)	0.91(0.82,1.00)
0.17	0.18	0.17(0.01,0.34)	0.17(-0.01,0.35)
0.18	0.16	0.15(-0.03,0.32)	0.15(-0.04,0.34)
0.11	0.14	0.15(-0.06,0.35)	0.14(-0.08,0.37)
0.14	0.11	0.10(-0.08,0.27)	0.10(-0.08,0.28)

4 IRT applications

In this section we illustrate the random loadings and cross-classified modeling framework with two practical IRT applications. In Section 4.1 we describe a multiple-group IRT model with measurement non-invariance and in Section 4.2 we illustrate a 2-parameter IRT model from the Generalizability theory point of view.

4.1 International Comparison of Student Achievement

In this section we describe another practical application of the random loading model using categorical data from the International Student Achievement (PISA) survey. The response data we analyze consists of 8 mathematics test items as in Section 7.6 in Fox (2010) where similar models are discussed. A total of 9796 students from 40 countries were sampled. The model we are interested is a simple 2-parameter IRT model that accommodates country non-invariance for all difficulty and discrimination parameters as well as factor means and variance. Let Y_{pij} denote the $p-$th item for individual i in country j. Consider the model

$$P(Y_{pij} = 1) = \Phi(a_{pj}\theta_{ij} - b_{pj}) \tag{17}$$

where θ_{ij} represents the individual math ability factor and is estimated as a standard normal random variable with mean 0 and variance 1. The function Φ is the standard normal distribution function. The variables a_{pj} and b_{pj} are the random discrimination and difficulty variables

$$a_{pj} = a_p + \xi_{a,pj} \tag{18}$$

$$b_{pj} = b_p + \xi_{b,pj}. \tag{19}$$

The parameters a_p and b_p are the mean discrimination and difficulty model parameters and $\xi_{a,pj}$ and $\xi_{b,pj}$ are the country specific deviations. The variables $\xi_{a,pj}$ and $\xi_{b,pj}$ are estimated as independent random effects with variance $\sigma^2_{a,p}$ and $\sigma^2_{b,p}$. This model allows us to explore a country specific 2-parameter IRT model. Let's call this model M1. The model has 32 parameters: 8 item discrimination and difficulties and their random effect variances. A natural extension to this model is to estimate a model that allows a cluster specific average of the math ability parameter

$$\theta_{ij} = \eta_j + \varepsilon_{ij} \tag{20}$$

where η_j is a country specific deviation and it is estimated as a random effect with mean zero and variance σ^2_η and ε_{ij} is standard normal variable with mean zero and variance 1. Let's denote this model as M2. Model M2 has just one more parameter than M1: σ^2_η. If the random discrimination parameters were actually fixed model M2 is essentially equivalent to a standard two-level factor analysis model where η_j is the factor on the between level and ε_{ij} is the factor on the within level and the factor loadings are the same on both levels. The interpretation of η_j in that case is that it is the common correlation effect for the difficulty random effects. It is not possible to estimate correlated difficulty random effects and country specific factor mean at the same time. Model M2 estimates independent random difficulty parameters and country specific factor mean.

A more general model that does not assume identical factor loadings for the within level and the between level can be estimated as well. This way the covariances between the random difficulty parameters would be estimated more accurately. This model is described by replacing equation (19) with the following equation

$$b_{pj} = b_p + \lambda_{b,p}\eta_j + \xi_{b,pj}. \tag{21}$$

where η_j is a standard normal random effect with mean 0 and variance 1. The variance is fixed to 1 for identification purposes because the $\lambda_{b,p}$ parameters are all free. This model essentially postulates that the variation in the average math ability between countries may have a different effect for the different items, just as the individual math ability may have a different effect on the different items. Lets denote this model by Model M3. The model has 40 parameters, the 8 loading parameters $\lambda_{b,p}$ and all of the parameters in model M1.

Model M4 extends the factor analysis model for the random difficulty parameters (21) to include also the random discrimination parameters. This way we can incorporate in the model correlation between the random difficulty and the random discrimination parameters. This is accomplished by replacing equation (18) with

$$a_{pj} = a_p + \lambda_{a,p}\eta_j + \xi_{a,pj}. \tag{22}$$

This model has an additional 8 parameters for a total of 48 parameters. If the number of items is more than half the number of countries the above model should be estimated using the BSEM methods described in Section 3.5 for identification purposes. In this particular example however the number of

countries is sufficiently large to estimate the model even without incorporating tiny priors for the loadings. Estimating all possible correlations between the random effects will yield more accurate estimates for the random effects and more accurate standard errors for the model parameters. When model M4 is estimated we find that first some of the loadings $\lambda_{a,p}$ and all of the loadings $\lambda_{b,p}$ are significant and thus the correlations between discrimination and difficulty parameters are significant. Therefore model M4 is a better fit for this data than model M3. All of the parameters $\lambda_{a,p}$ are relatively small and thus the interpretation of η_j as the country specific factor mean is still preserved however if in another application the loadings $\lambda_{a,p}$ are large and $\lambda_{b,p}$ are small the interpretation of η_j should be different.

Model M5 is a model where we allow country specific factor variance as well. Instead of having the variance of ε_{ij} as fixed to 1 we estimate the following equation

$$Var(\varepsilon_{ij}) = 0.51 + (0.7 + \zeta_j)^2 \tag{23}$$

where ζ_j is a country specific random effect with mean 0 and variance σ_v^2. If the variance parameter σ_v^2 is zero than $Var(\varepsilon_{ij}) = 1$. The variation in the country specific factor variance is determined by the magnitude of σ_v^2. The above equation is somewhat arbitrary in that it essentially limits the variation in the country specific factor variance so that the minimal variance in a country is 51% of the average. However this is not really a limitation because in practical applications this limit will not be reached and a different equation can be estimated instead. For example to allow more variation in the factor variance one can estimate

$$Var(\varepsilon_{ij}) = 0.1 + (0.95 + \zeta_j)^2 \tag{24}$$

or even

$$Var(\varepsilon_{ij}) = (1 + \zeta_j)^2. \tag{25}$$

For this application however we use equation (23). Model M5 just one more parameter σ_v^2 for total of 49 parameters.

After estimating model M5 and testing the proportionality of the between and within factor loadings we conclude that the loadings are proportional in this example. Thus we estimate model M6 which is the same as M5 but the loading parameters on the within and the between level are the same, i.e., $\lambda_{b,p} = a_p$. In this model the variance of η_j is estimated as a free parameter. Model M6 has 42 parameters. The loading proportionality is

tested simply by evaluating the distribution of new derived parameters. For example to evaluate the proportionality of the loadings for the first and the second item we evaluate the distribution of $\lambda_{b,1}a_2 - \lambda_{b,2}a_1$. If this parameter is not significantly different from 0 then we conclude that the loadings for the first two items are proportional.

There are three main differences between the models described above and the most general model described in Section 7.6 in Fox (2010). First, equation (21) gives a more flexible model because the country specific variation in the math ability can have a different effect for each item. Second, incorporating the random difficulty parameters in the between level factor model allows us to estimate correlations between all random difficulty and random discrimination parameters. Third, the model identification is accomplished simply by restricting parameters rather than imposing restrictions on the random effects. For example the mean of η_j is fixed to 0. Such an approach is easy to interpret and understand. The most general model in Section 7.6 in Fox (2010) is essentially equivalent to model M2 with the addition of equation (23).

The parameter estimates of model M6 are presented in Table 13 and 14. The discrimination parameters are held equal to the difficulty loading parameters and that can be seen in the tables. The two other model parameters are $SD(\eta_j)$ and $SD(\zeta_j)$ are estimated as 0.497(0.065) and 0.063(0.034). The fact that $SD(\zeta_j)$ is marginally significant indicates that the factor variance is nearly invariant across countries. It is worth noting that the first two discrimination loading parameters are significant and thus they improve the model fit. When using random effects it is important to be able to evaluate the significance of the variance component as this establishes the need to use random parameters rather than fixed parameters. An approximate way to do this is just to use maximum likelihood style t-test using the standard errors reported from the Bayes estimator. The logic behind this approach is justified by the fact that asymptotically as the number of clusters increases the Bayes estimates and their standard errors are equivalent to the ML estimates and thus using a simple t-value approach is not unfounded especially when the number of clusters is large. Using this method we see that all difficulty variance parameters are easily significant while some discrimination variance are significant and some are borderline significant, namely, those for items 1, 3 and 7. A more advanced approach for testing variance significance for random effects has been described in Verhagen and Fox (2012) based on Bayes factor methodology. Using variance components prior of $IG(1, 0.005)$

and testing the model that the variance components is less than 0.001 we find the same results, i.e., the variance components for all difficulty parameters are significant while some of the discrimination variance components are significant and some are not significant. The Bayes factor for the discrimination variance being zero for item 3 is 3.1 and for the rest of the parameters it is less than 3. When the Bayes factor exceeds the threshold value of 3 we interpret that as a substantial evidence in favor of the tested hypothesis.

The variance component for the ζ_j which models country specific factor variance is also borderline significant with the Bayes factor for the variance component being 1.9. The Bayesian methodology is flexible and the random effects that are not significant can simply be eliminated from the model or they can be estimated even though there is little evidence that they are needed. The harm of estimating a random effect with insignificant variance component is negligible.

Finally using Model M6 we can estimate all latent variables and cluster specific random effect. Table 15 contains the estimates for the cluster specific math ability mean parameter η_j and its confidence limits based on 30 distribution draws. We report the parameter values only for the top 3 and the bottom 3 countries. One can also obtain the entire distribution for a specific random effect. For example in Figure 1 we display the country specific distribution for the mean ability parameter for Finland using 300 distribution draws.

The ranking in Table 15 is of course not statistically significant. For example the confidence interval for FIN and KOR substantially overlap. A proper comparison can be conducted by evaluating the posterior distribution for the difference between the two random effect which can easily be obtained from the posterior distribution of the two random effects. In general, such a comparison can be done also within the maximum-likelihood estimation however with the Bayesian estimation the computation is more accurate because it accounts for the posterior correlation between these random effects that is due to their dependence on the common model parameters.

Table 13: Discrimination parameters and standard errors for the 2-parameter IRT model

Item	a_p	$\sigma_{a,p}$	$\lambda_{a,p}$
1	0.732 (0.031)	0.065 (0.036)	-0.148 (0.074)
2	1.032 (0.049)	0.180 (0.056)	0.373 (0.108)
3	0.635 (0.028)	0.066 (0.028)	-0.027 (0.063)
4	0.624 (0.031)	0.083 (0.033)	0.101 (0.062)
5	0.528 (0.027)	0.111 (0.034)	-0.009 (0.062)
6	0.347 (0.031)	0.144 (0.049)	-0.046 (0.075)
7	0.625 (0.030)	0.058 (0.034)	-0.024 (0.073)
8	0.619 (0.029)	0.092 (0.040)	-0.103 (0.094)

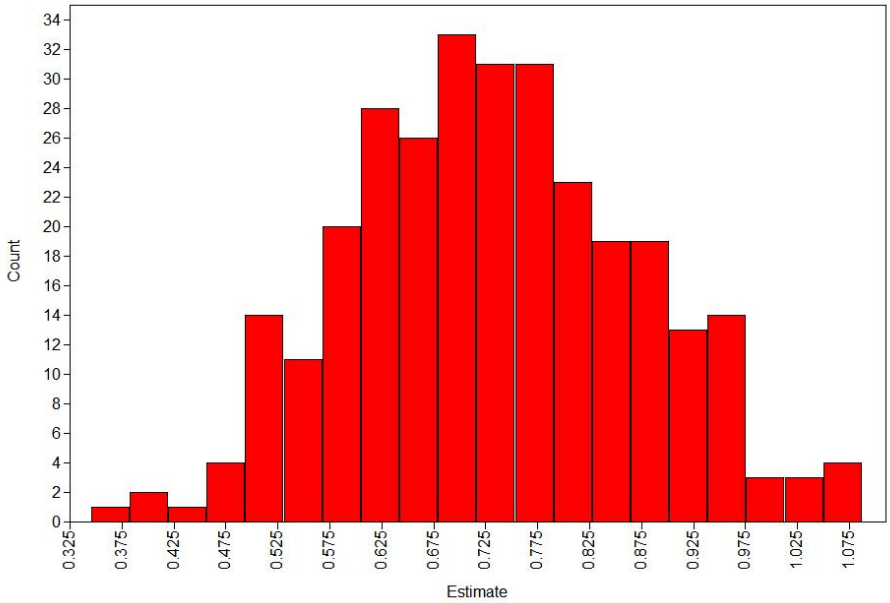
Table 14: Difficulty parameters and standard errors for the 2-parameter IRT model

Item	b_p	$\sigma_{b,p}$	$\lambda_{b,p}$
1	-0.581 (0.053)	0.236 (0.038)	0.732 (0.031)
2	0.121 (0.070)	0.265 (0.045)	1.032 (0.049)
3	-0.056 (0.047)	0.189 (0.034)	0.635 (0.028)
4	-0.405 (0.050)	0.189 (0.034)	0.624 (0.031)
5	-0.037 (0.042)	0.130 (0.025)	0.528 (0.027)
6	-1.554 (0.038)	0.194 (0.039)	0.347 (0.031)
7	-0.808 (0.057)	0.171 (0.033)	0.625 (0.030)
8	-0.969 (0.047)	0.093 (0.033)	0.619 (0.029)

Table 15: Country specific mean ability parameter

Country	Estimate and confidence limits
FIN	0.749 (0.384 , 0.954)
KOR	0.672 (0.360 , 0.863)
MAC	0.616 (0.267 , 1.041)
BRA	-0.917 (-1.166 , -0.701)
IDN	-1.114 (-1.477 , -0.912)
TUN	-1.156 (-1.533 , -0.971)

Figure 1: Country specific distribution for the mean ability parameter for FIN



4.2 Random Item Example: Generalizability Theory

Generalizability theory postulates that test items can be considered a random samples from a population of items, see De Boeck (2008). This concept is particularly useful when the same or different items are administered to different individuals and is typically used to model computer generated items or items that are used in adaptive testing. Let Y_{ij} denote the binary outcome from a test item j administered to individual i. Consider the 2 parameter random IRT model

$$P(Y_{ij} = 1) = \Phi(a_j\theta_i - b_j) \tag{26}$$

where all there variables are random effects. The random discrimination effect a_j is normally distributed with mean a and variance σ_a^2 and the random difficulty parameter is normally distributed with mean b and variance σ_b^2. The ability variable θ_i is normally distributed with mean zero and variance 1. The above model is a cross-classified model where the two random modes are individual and item. Each cross classified cell has 1 or 0 observations. The latent variable a_j is the random loading for the ability factor θ_i and thus this model is a special case of the general random loading cross-classified SEM framework described above. The model has only 4 parameters and thus it is much more parsimonious than the regular IRT model which has as many parameters as two times the number of items. To illustrate the above model we use the data from the Trends in International Mathematics and Science Study (TIMMS) 2007 used in Fox (2010) Section 4.3.3. The data consists of 8 math items administered to 478 fourth graders. The model parameters are presented in Table 16.

In this example the number of random items is only 8, i.e., there are only 8 clusters in the random dimensions of items. In such a situation the estimates for the variance parameters of the random effects will be sensitive to the prior specification. Using a weakly informative prior is generally preferred. The above estimates are obtained with the following weakly informative priors. The prior for σ_a^2 is IG(1,0.1) and the prior for σ_b^2 is IG(1,1).

Using the posterior distribution of the random effects we can also estimate the item specific discrimination and difficulty parameters. To estimate the posterior mean and standard errors we use 300 draws from the posterior distribution. These estimates are presented in Table 17. In Table 18 we compare the random 2-parameter IRT model estimates for the discrimination parameter and the the discrimination parameter estimates obtained with the standard IRT model where all item parameters are estimated as

Table 16: Random 2-parameter IRT

parameter	estimate	SE
a	0.752	0.094
b	0.118	0.376
σ_a^2	0.050	0.046
σ_b^2	1.030	0.760

fixed parameter using the ML estimator. We can see that the Bayes random estimates are shrunk towards the mean and have smaller standard errors which suggests that these estimates are more accurate because the standard errors are a proxy for the square root of the mean squared error. The ML standard errors are bigger on average by 13%. To verify that this interpretation is correct we generate data using (26) and the parameters values in Table 16 using a data set with 500 observations and 10 items. We then estimate model (26) on this data set as well as the 2-parameter ML-based IRT model. The results from the simulation study for the discrimination parameters are presented in Table 19. The ML standard errors are bigger again by 14% on average and the mean squared error is bigger by 19%, i.e., we see that indeed the observed decrease in standard errors obtained by the Bayes estimator is a proxy for more accurate estimates. The actual values match as well. For the Bayes estimator the average standard errors is 0.101 and the square root of the mean squared error is 0.092 while for the ML estimator these values are 0.115 and 0.110. In this generated data set the accuracy gain in the estimation of the parameter estimates obtained by the Bayes estimator did not result in a substantial decrease in the mean squared error estimation of the ability parameter. The MSE for the Bayes estimator is 0.574 while for the ML estimator it is 0.575. The means squared difference between the Bayes ability estimate and the ML ability estimate is 0.043, i.e., the two estimates are very close and the substantial portion of the error in the estimation of the ability parameters is due to measurement error and can not be eliminated by either estimator. In the simulated data analysis using weakly informative priors or non-informative priors did not affect the mean squared error comparisons.

Table 17: Random 2-parameter IRT item specific parameters

item	discrimination	SE	difficulty	SE
Item 1	0.797	0.110	-1.018	0.103
Item 2	0.613	0.106	-0.468	0.074
Item 3	0.905	0.148	-1.012	0.097
Item 4	0.798	0.118	-1.312	0.106
Item 5	0.538	0.099	0.644	0.064
Item 6	0.808	0.135	0.023	0.077
Item 7	0.915	0.157	0.929	0.090
Item 8	0.689	0.105	1.381	0.108

Table 18: Random 2-parameter IRT example comparison with standard 2-parameter IRT model based on the ML estimator

item	Bayes random discrimination	Bayes random SE	ML fixed discrimination	ML fixed SE
Item 1	0.797	0.110	0.850	0.155
Item 2	0.613	0.106	0.579	0.102
Item 3	0.905	0.148	0.959	0.170
Item 4	0.798	0.118	0.858	0.172
Item 5	0.538	0.099	0.487	0.096
Item 6	0.808	0.135	0.749	0.119
Item 7	0.915	0.157	0.929	0.159
Item 8	0.689	0.105	0.662	0.134

Table 19: Simulation study results: Random 2-parameter IRT comparison with standard ML based 2-parameter IRT model

item	True Value	Bayes random discrimination	Bayes random SE	ML fixed discrimination	ML fixed SE
Item 1	0.608	0.642	0.088	0.606	0.094
Item 2	0.875	0.853	0.106	0.870	0.128
Item 3	0.520	0.650	0.089	0.619	0.095
Item 4	0.764	0.766	0.098	0.751	0.114
Item 5	0.834	0.985	0.118	1.072	0.148
Item 6	0.623	0.506	0.087	0.446	0.093
Item 7	0.876	0.858	0.116	0.889	0.127
Item 8	0.776	0.797	0.099	0.794	0.111
Item 9	0.581	0.543	0.095	0.497	0.100
Item 10	0.979	0.812	0.119	0.853	0.144

5 Individual Differences Factor Analysis

The example in section 4.1 provides a factor analysis modeling framework where variability in the factor variance across groups can be estimated simultaneously with the variability in the factor loading. That parameterization however is prone to slow convergence or poor mixing. In addition that model has a somewhat deficient connection between the factor variance variability and the factor loading variability. Here we provide a different model that can be used to more clearly separate the variability of the factor model across groups and the variability in the factor model.

To illustrate this model we will use an example presented in Jahng et al. (2008). [1] The data is obtained from an ongoing study of affective instability in borderline personality disorder (BPD) patients. Affective instability is considered a core feature of BPD that distinguishes this disorder from other disorders like depressive disorders. The data contains 84 individuals. Two groups of outpatients were entered into the study, 46 individuals with borderline personality disorder and 38 with major depressive disorder or dys-

[1] We thank Tim Trull and Phil Wood for providing the data and providing helpful comments

thymic disorder. The mood factor for each individual was measured with 21 self-rated items on a scale of 1-5. We analyze the data assuming continuous distributions for the items. For each individual the measurements were collected randomly several times a day over a 4 week period. In total between 76 to 186 assessments were conducted per person.

An analysis of the 21-item measurement instrument needs to take into account that repeated observations over time are correlated within individuals. One possible analysis approach is two-level factor analysis where repeated measures are nested within individuals. Let Y_{pij} be the p-th item for individual i at assessment j and X_i be the binary indicator for the individual being classified with borderline personality disorder. Let P be the number of assessment items. A two-level factor model can be defined by the following two equations

$$Y_{pij} = \mu_p + \lambda_p \eta_{ij} + \zeta_{pi} + \varepsilon_{pij} \tag{27}$$

$$\eta_{ij} = \eta_i + \beta X_i + \xi_{ij}. \tag{28}$$

Here μ_p is the intercept parameter for the p-th item, ζ_{pi} is the individual variation from this mean parameter. The mean of ζ_{pi} is fixed to 0 for identification purposes and the variance parameter v_p is estimated. The factor η_{ij} is decomposed as the sum of the mean factor value η_i for individual i and the assessment specific deviation from that mean ξ_{ij}. The loading parameters λ_p are all estimated as well as the variance ψ of η_i while the variance of ξ_{ij} is fixed to 1 for identification purposes. The residual ε_{pij} is a zero mean normally distributed residual with variance θ_p. This model has $4P$ parameters: μ_p, λ_p, θ_p, and v_p as well as two parameters ψ and β for a total of $4P + 2$ parameters.

The two-level factor model, however, has important shortcomings for these types of data. The model accommodates individually specific random intercepts for the factor as well as for each factor indicator through the latent variables η_i and ζ_{pi}. However, the model does not accommodate individual-specific factor variances, nor individual-specific factor loadings. The individual-specific factor variance is a key indicator of the individual's stability over time, in this case mood stability. Across-individual differences in these factor variances is informative about individual differences in the factor. Individual-specific factor loadings capture possibly different reactions to the various items for different individuals. Different individuals may use a certain item's scale differently. The study of factor loading variation across individuals therefore informs about which items are most suitable for

comparing individuals. A model is needed that can naturally separate the across-individual variation in the factor loadings and the across-individual variation in the factor variance. A new model is introduced here which allows for this flexibility. The fact that many repeated measures are observed for each individual makes for reliable estimation of these quantities.

Note again here that in our example the individual variable takes the role of the grouping variable because there are multiple assessments for each individual. To accommodate individual-specific factor variance we can introduce an individual-specific scaling factor σ_i in equation (27)

$$Y_{pij} = \mu_p + \lambda_p \sigma_i \eta_{ij} + \zeta_{pi} + \varepsilon_{pij}. \tag{29}$$

Thus the total factor $\sigma_i \eta_{ij}$, conditional on i, has individual specific variance σ_i^2. The random effect variable σ_i is defined on the individual/between level as a normally distributed random variable. The mean of σ_i is fixed to 1 for identification purposes and the variance σ^2 is estimated. A different interpretation of the above model is that we have individually specific factor loadings s_{pi}

$$s_{pi} = \lambda_p \sigma_i. \tag{30}$$

In such a model however the correlations between the individual specific factor loadings is 1 because the loadings are all proportional to the random effect σ_i. To allow individual variation in the loadings we can introduce a residual in equation (30)

$$s_{pi} = \lambda_p \sigma_i + \epsilon_{pi} \tag{31}$$

where ϵ_{pi} has a zero mean and variance w_p. In this model we can clearly see the separation between the variation in the factor variance and the variation in the factor loadings. The term $\lambda_p \sigma_i$ in equation (31) corresponds to the variation in the factor variance, while the term ϵ_{pi} corresponds to the variation in the factor loadings that goes beyond what can be explained by the variation in the factor variance.

The above model essentially estimates random loadings for the basic factor analysis model (27) and then on the between level it estimates a factor analysis model for the random factor loadings. Factor analysis estimation tends to absorb most of the correlation between the indicators within the factor model and to minimize the residual variances. This corresponds to what we want in this two level model. The factor analysis on the factor loadings will tend to absorb as much as it can in the factor analysis portion

of the model and minimize the residual variances. Thus the model will try to explain as much as possible the variation between the correlation matrices across individual as a variation in the factor variance rather than as a variation in the factor loadings. Thus this model is ideal for evaluating and separating the loading non-invariance and the factor variance non-invariance. Testing $w_p = 0$ is essentially a test for measurement invariance. Testing $\sigma^2 = 0$ is essentially a test for factor variance invariance across the groups. Testing $\psi = 0$ is essentially a test for factor mean invariance across the groups.

The final model we estimate on the borderline personality disorder, combines the above equations and is described as follows

$$Y_{pij} = \mu_p + s_{pi}\eta_{ij} + \zeta_{pi} + \varepsilon_{pij} \tag{32}$$

$$\eta_{ij} = \eta_i + \beta_1 X_i + \xi_{ij} \tag{33}$$

$$s_{pi} = \lambda_p \sigma_i + \epsilon_{pi} \tag{34}$$

$$\sigma_i = 1 + \beta_2 X_i + \zeta_i. \tag{35}$$

Here we regress the scale factor σ_i on the BPD covariate X_i. The regression coefficient β_2 shows the amount of increase or decrease of the factor variance due to the borderline personality disorder, i.e., the two regression coefficients β_1 and β_2 represent the effect of the covariate X_i on the mean and the variance of the mood factor. The above model has an additional $P + 2$ parameters compared to model (27-28). These are the parameters β_2, $Var(\zeta_i) = \sigma^2$ and the P parameters w_p. An analysis based on this model will be called an Individual Differences Factor Analysis (IDFA). To summarize, the model accommodates individually specific factor intercept and variance, individually specific factor loadings as well as individually specific intercepts for each indicator variable.

It should be noted that a consequence of using IDFA is that individuals can be compared on their factor scores even when there is variation in both the factor variances and factor loadings. The scores for the individual-level factor component η_i in (33) can be estimated using Bayes plausible values. This implies that individuals can be compared despite a certain amount of measurement non-invariance in the factor loadings.

The parameter estimates and standard errors for the IDFA model are presented in Table 20. There are 4 parameters not presented in Table 20: β_1, β_2, ψ and σ^2 and the estimates for these parameters are as follows: 0.613(0.119),

0.087(0.128), 0.462(0.091), 0.268(0.066) respectively. It is interesting to note that β_2 is not statistically significant, i.e., the individuals with borderline personality disorder do not have significantly larger mood variation compared to individuals with major depressive disorder.

In Table 20 we also present the percentage of the factor loading variation that can be explained by the variation in the factor variance. That percentage varies from 24% to 80%. Testing the statistical significance for w_p shows that all variance components are statistically significant, i.e., the loadings parameters should be individually specific. This is a clear evidence that measurement instruments may not be interpreted the same way by different individuals and thus individual specific adjustments are needed to properly measure underlying factors. This is probably true even for many cross-sectional studies however individual specific adjustments to the loadings in factor analysis can be made only when repeated assessments are conducted.

Table 20: Differential Factor Analysis: Comparing loading non-invariance and factor variance non-invariance in borderline personality disorder study.

item	θ_p	λ_p	μ_p	v_p	w_p	Percentage loading invariance
Item 1	0.444(0.006)	0.261(0.036)	1.505(0.060)	0.287(0.049)	0.045(0.009)	0.29
Item 2	0.628(0.008)	0.377(0.049)	1.524(0.080)	0.482(0.081)	0.080(0.014)	0.32
Item 3	0.331(0.004)	0.556(0.062)	1.209(0.046)	0.057(0.010)	0.025(0.006)	0.77
Item 4	0.343(0.005)	0.553(0.063)	1.301(0.053)	0.097(0.018)	0.030(0.006)	0.73
Item 5	0.304(0.004)	0.483(0.057)	1.094(0.031)	0.017(0.004)	0.053(0.010)	0.54
Item 6	0.477(0.007)	0.471(0.056)	1.489(0.056)	0.188(0.032)	0.055(0.011)	0.52
Item 7	0.328(0.005)	0.506(0.058)	1.450(0.065)	0.255(0.042)	0.025(0.005)	0.73
Item 8	0.297(0.004)	0.516(0.059)	1.312(0.061)	0.207(0.036)	0.018(0.004)	0.80
Item 9	0.313(0.004)	0.386(0.049)	1.207(0.037)	0.083(0.015)	0.049(0.009)	0.45
Item 10	0.213(0.003)	0.401(0.054)	1.147(0.041)	0.103(0.019)	0.066(0.011)	0.40
Item 11	0.301(0.004)	0.495(0.057)	1.412(0.064)	0.272(0.045)	0.024(0.005)	0.73
Item 12	0.291(0.004)	0.402(0.050)	1.168(0.035)	0.072(0.014)	0.048(0.009)	0.47
Item 13	0.466(0.006)	0.443(0.054)	1.638(0.068)	0.328(0.056)	0.041(0.008)	0.56
Item 14	0.167(0.002)	0.339(0.051)	1.110(0.040)	0.091(0.016)	0.081(0.014)	0.28
Item 15	0.511(0.007)	0.471(0.058)	1.537(0.085)	0.528(0.086)	0.077(0.013)	0.44
Item 16	0.382(0.005)	0.330(0.042)	1.517(0.076)	0.430(0.073)	0.044(0.008)	0.40
Item 17	0.185(0.003)	0.386(0.051)	1.150(0.047)	0.130(0.023)	0.066(0.011)	0.38
Item 18	0.187(0.003)	0.325(0.050)	1.015(0.016)	0.002(0.001)	0.092(0.016)	0.24
Item 19	0.131(0.002)	0.354(0.053)	0.982(0.016)	0.007(0.002)	0.105(0.018)	0.24
Item 20	0.269(0.004)	0.476(0.060)	1.101(0.033)	0.038(0.008)	0.095(0.016)	0.39
Item 21	0.167(0.002)	0.379(0.051)	1.004(0.018)	0.006(0.002)	0.096(0.017)	0.29

6 Intensive Longitudinal Data

Intensive longitudinal data have become quite common in recent years due to accumulating long-term longitudinal studies as well as a result of the use of new technological devices for data collection such as mobile devices, beepers and web interfaces, see Walls and Schafer (2006). New models need to be developed to analyze such data. Longitudinal data typically has been analyzed with multivariate models however if the number of time points is large these models can fail due to too many variables and parameters involved in the modeling. Estimating structural latent variable models in intense longitudinal settings can lead to additional challenges. Factor analysis models may be unstable over time and measurement invariance may be violated to some degree. Thus the time invariant structural models would be insufficient and inaccurate. The framework described in this article can resolve these problems. The random loading and intercept models can be used to model measurement and intercept non-invariance. These models have the advantage of borrowing information over time in the estimation the same way standard structural models do when assuming invariance. At the same time these new models have the advantage of accommodating measurement non-invariance the same way longitudinal structural models do. The models are also more parsimonious than longitudinal structural models. All these advantages will typically lead to more accurate model estimation.

To illustrate the intensive longitudinal modeling we will use the TOCA example described in Ialongo et al. (1999). The data consists of a teacher-rated measurement instrument capturing aggressive-disruptive behavior among a sample of U.S. students in Baltimore public schools. The instrument consists of 9 items scored as 0 (almost never) through 6 (almost always). A total of 1174 students are observed in 41 classrooms from Fall of Grade 1 through Grade 6 for a total of 8 time points. The multilevel (classroom) nature of the data is ignored in the current analysis. The item distribution is very skewed with a high percentage in the Almost Never category. The items are therefore dichotomized into the Almost Never versus all the other categories combined. For each student a 1-factor analysis model is estimated with the 9 items at each time point.

In the following sections we illustrate the three different approaches to intensive longitudinal data modeling: Longitudinal SEM, Multilevel SEM, Cross-classified SEM. We discuss the advantages and disadvantages of each method.

6.1 Longitudinal SEM

Let Y_{pit} be the $p-$th item for individual i at time t. The factor analysis model at time t is described by

$$P(Y_{pit} = 1) = \Phi(\lambda_{pt}\eta_{it} - \tau_{pt}). \tag{36}$$

The variance of the aggressive behavior factor η_{it} is fixed to 1 and the mean if fixed to 0 for identification purposes. There are 8 times points and 9 items so in total we have 72 thresholds parameter τ_{pt} and 72 loading parameters λ_{pt}. In addition, the aggressive behavior factors η_{it} are nested within individual and thus we should account for the correlation between the individual factors across time. One approach is to simply estimate an unrestricted correlation matrix for the 8 individual factors. This would contribute an additional 28 correlation parameters for a total of 172 parameters. Because this model has 8 latent variables it cannot be estimated easily with the ML estimator which would requires 8 dimensional integration. The model can be estimated however with the WLSMV estimator as implemented in Mplus program, see Muthén and Muthén (1998-2010). The main problem with this model however is that it is not scalable in terms of time. The number of parameters grows as a quadratic function of the number of time points. In addition, the model is estimated as a multivariate model and this estimation is based on fitting an unrestricted multivariate probit model with 72 variables. This model has 72 threshold parameters and 2556 correlation parameters. The estimation of the unrestricted model can easily become computationally prohibitive as the number of time points increases. In addition, the sample size needed to estimate this unrestricted model may be substantial.

While the above model is flexible and accounts for measurement and threshold non-invariance it is difficult to provide interpretation for the variation in the parameters across time and to guarantee that the factor measurement model is sufficiently stable so that we can interpret it as the same factor changing over time. This leads to estimating additional models with ad-hoc parameter restrictions designed to parse those model parameters that are significantly different from those that are not. This process however is not feasible when the number of time points is substantial. A reasonable model to explore is the model that assumes complete measurement and threshold invariance

$$P(Y_{pit} = 1) = \Phi(\lambda_{p}\eta_{it} - \tau_{p}). \tag{37}$$

In addition we can impose a linear or quadratic growth model for the aggressive behavior factors η_{it}. The linear trend model is described as follows

$$\eta_{it} = \mu_t + \alpha_i + \beta_i \cdot t + \varepsilon_{it}. \tag{38}$$

The variables α_i and β_i are individual level random effect, ε_{it} is the residual of the growth model and μ_t is a time specific parameter. The variance of ε_{it} is also a time specific parameter θ_t. The means of α_i and β_i are fixed to 0 for identification purposes. The intercept and residual variance parameters at time $t = 0$, μ_0 and θ_0, are also fixed to 0 and 1 for identification purposes. The above model can be estimated with the WLSMV estimator and it has 35 parameter only: 9 threshold and loading parameters, 7 factor intercept and factor residual variances as well as the three parameters in the variance covariance matrix of α_i and β_i. The growth model here is a model for the individual growth of the aggressive behavior factor and it reflects the factor variation beyond the average change over time which is modeled with the parameters μ_t. If the linear trend model in (38) does not hold, the β_i random effect will be estimated to 0 and its variance will be estimated to zero as well. An alternative model is the model where the linear change over time includes modeling the change in the factor mean. This can be achieve by removing the parameters μ_t and estimating a mean parameter for β_i. However, often in practical applications the parameters μ_t will not follow a linear trend and such a model would most likely lead to a model misfit. The above model can only be estimated as a multivariate model with the WLSMV estimator. The model is more parsimonious than model (36) however relies on the assumption of measurement and intercept invariance. The model is also not scalable with respect to time.

6.2 Multilevel SEM

A different approach to modeling time intensive data is two-level modeling where the cluster variable is the individual and the observations at the different time points are the observations within cluster. Consider for example the model

$$P(Y_{pit} = 1) = \Phi(\lambda_p \eta_{it} - \tau_p) \tag{39}$$

$$\eta_{it} = \alpha_i + \beta_i \cdot t + \varepsilon_{i.t} \tag{40}$$

The variance of ε_{it} is fixed to 1 for identification purposes and the random effects α_i and β_i are individual level growth factors. The random effect α_i

has a mean 0 for identification purposes. The mean of β_i is not zero, i.e., in this model we actually restricted the parameters μ_t from (38) to follow a linear trend. This model is time scalable. The number of parameters is 21: 9 threshold and loading parameters as well as the mean of β_i and the variance parameters of α_i and β_i. This model can be estimated as a two-level model with 3 dimensional numerical integration and it can be estimated with a large number of time points without any computational problems. In fact, the more time points there are the more accurate the estimates of the random effects and the model parameters. Thus this model has the advantage of being time scalable. However, the model has the disadvantage that it assumes loading invariance and the threshold non-invariance is limited.

Another advantage of two-level models for intensive longitudinal data is the fact that we can accommodate more individual level random effects. Consider for example the model

$$P(Y_{pit} = 1) = \Phi(\lambda_p \eta_{it} - \tau_{pi}). \tag{41}$$

$$\eta_{it} = \alpha_i + \varepsilon_{it} \tag{42}$$

$$\tau_{pi} = \tau_p + \epsilon_{pi} \tag{43}$$

The difference between this model and the model (39-40) is that it accommodates individual level variation ϵ_{ip} in the threshold parameters. Such random effects are useful in modeling individual level variation from the factor model. For example, individuals may score lower or higher on a particular item, inconsistent with the rest of the items, due to reasons other than the aggressive behaviors factor. This model has 28 parameters: 9 thresholds, loadings and individual level residual variances as well as the variance parameter of α_i. This model is also scalable in terms of time and it can be estimated with the WLSMV estimator as a two-level model. The model cannot be estimated easily with the ML estimator because such an estimation would require 10 dimensional numerical integration. The Bayes estimator can also be used to estimate this model, in fact, the Bayes estimator can be used to estimate a model that combines all of the features in models (39-40) and (41-43), including the linear trend random effect β_i.

In the above two-level models the latent variable α_i is essentially a between level factor where the between level factor loadings are the same as the within level factor loadings. This however need not be the case. Separate loadings can be estimated for η_{it} and α_i. Other variations of two-level models are possible as well. All two-level models however are based on assuming

measurement and threshold invariance across time. It is possible to introduce dummy variable predictors for the different time points and thus accommodate threshold non-invariance however such models are not time-scalable due to the many parameters and dummy variables in the model.

6.3 Cross-classified SEM

The cross-classified SEM framework described in this article which accommodates cross-classified modeling as well as random loadings can be used to construct time scalable models that can also accommodate non-invariance for the loadings and threshold parameters. Growth modeling for the factor variable can also be estimated. We use the cross-classified SEM framework because in the intensive longitudinal data the observations are cross-nested within individual and time. Consider the model

$$P(Y_{pit} = 1) = \Phi(\lambda_p \eta_{it} - \tau_{pit}). \tag{44}$$

$$\eta_{it} = \alpha_i + \varepsilon_{it} \tag{45}$$

$$\tau_{pit} = \tau_p + \epsilon_{pi} + \zeta_{pt} \tag{46}$$

This model is very similar to model (41-43). It has an additional random effect ζ_{pt} which has 0 mean and variance ψ_p. These random effects accommodate threshold non-invariance across time. The model has a total of 37 parameters, all of the parameters in model (41-43) plus the 9 random effect variance parameters ψ_p. The next model we consider is the model that accommodates the features of models (44-46), (39-40) and (41-43). This is accomplished by substituting equation (45) with equation (40). This model has just one more parameter: the variance of β_i for a total of 38 parameters. The next modeling extension is to accommodate measurement non-invariance across time. Consider the model

$$P(Y_{pit} = 1) = \Phi(\lambda_{pt} \eta_{it} - \tau_{pit}). \tag{47}$$

$$\eta_{it} = \alpha_i + \beta_i \cdot t + \varepsilon_{it} \tag{48}$$

$$\tau_{pit} = \tau_p + \epsilon_{pi} + \zeta_{pt} \tag{49}$$

$$\lambda_{pt} = \lambda_p + \xi_{pt}. \tag{50}$$

The random effects ξ_{pt} allow variation across time in the measurement model. The mean of ξ_{pt} is 0 and the variance is w_p. This model has just 9 more parameter, the variances parameters w_p for a total of 47 parameters. The final

modification we make to accommodate all features of the previous models while retaining the time scalability and measurement and threshold non-invariance is the possibility to have the variance of ε_{it} vary across time as in model (37-38). This is accomplished as in equation (23). We estimate the model where

$$Var(\varepsilon_{it}) = 0.51 + (0.7 + \gamma_t)^2. \tag{51}$$

The random effect γ_t has a mean of 0 and variance σ^2. Note here that we could also estimate a model where the aggressive behavior factor has a time non-invariant mean however such a random effect would not be identified simultaneously with the random effects ζ_{pt} that allow the time non-invariance for each item rather than only for the factor. We estimate model (47-51) and present the results in Table 21. The model has a total of 48 parameters. In table 21 we report the point estimates and standard errors for the item specific parameters. The estimates for the variance of the random effects for α_i, β_i and γ_t are 1.98(0.17), 0.10(0.01) and 1.44(1.86). For all variance parameters the standard errors should not be used for testing significance. Instead performing a test of significance as in Verhagen and Fox (2012) yields that all variance components are significant with the exception of the random effect for ζ_{8t}. For this random effect the Bayes factor for the hypothesis $Var(\zeta_{8t}) < 0.001$ is 2, i.e., the variance component is marginally significant.

The results presented in Table 21 are obtained with uninformative priors for all parameters. In this example there are only 8 time points. Thus the results for the time specific random effects ζ_{pt} and ξ_{pt} will be sensitive to the prior specification of the variance parameters. Weakly informative priors rather than uninformative prior would be more appropriate to use.

Table 21: Cross-classified intensive longitudinal growth model: TOCA example

item	τ_p	Variation across time $Var(\tau_{pit}\|i) = Var(\zeta_{pt})$	Variation across individual $Var(\tau_{pit}\|t) = Var(\epsilon_{pi})$	λ_p	Variation across time $Var(\lambda_{pt}) = Var(\xi_{pt})$
Item 1	-0.87(0.15)	0.12(0.19)	0.15(0.04)	0.70(0.08)	0.03(0.07)
Item 2	-0.83(0.12)	0.05(0.10)	0.16(0.05)	0.97(0.13)	0.09(0.14)
Item 3	1.00(0.16)	0.11(0.23)	0.02(0.05)	1.31(0.21)	0.22(0.41)
Item 4	1.45(0.19)	0.17(0.33)	0.10(0.05)	0.97(0.16)	0.12(0.20)
Item 5	-0.09(0.12)	0.06(0.11)	0.24(0.04)	0.90(0.08)	0.03(0.05)
Item 6	0.90(0.18)	0.16(0.30)	0.08(0.03)	1.03(0.15)	0.10(0.17)
Item 7	0.28(0.14)	0.09(0.18)	0.16(0.04)	1.01(0.11)	0.06(0.12)
Item 8	0.15(0.07)	0.01(0.08)	0.13(0.04)	0.91(0.11)	0.06(0.10)
Item 9	-0.51(0.13)	0.08(0.13)	0.20(0.04)	0.94(0.12)	0.07(0.14)

7 Conclusion

The Bayesian estimation of structural equation models has become more popular as stable numerical algorithms have been developed. It is now possible to explore models that go beyond the reach of traditional ML and WLS estimators using the Bayesian estimation. Cross classified structural models and random loading models are two such examples. Using these new models it is now possible to address data modeling issues that were not possible to address within the standard structural modeling framework.

References

[1] Asparouhov, T. & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. Proceedings of the Joint Statistical Meeting, 2531-2535.

[2] Asparouhov, T. and Muthén, B. (2009). Exploratory structural equation modeling. Structural Equation Modeling, 16, 397-438.

[3] Asparouhov, T. and Muthén, B. (2010a). Bayesian analysis using Mplus: Technical implementation. Technical Report.
http://statmodel.com/download/Bayes3.pdf

[4] Asparouhov T. and Muthén B. (2010b) Bayesian Analysis of Latent Variable Models using Mplus. Technical Report.
http://statmodel.com/download/BayesAdvantages18.pdf

[5] Asparouhov T. and Muthén B. (2012) Comparison of computational methods for high dimensional item factor analysis. Mplus Ttechnical Report.

[6] Brennan, R. L. (2001). Generalizability Theory. New York: Springer-Verlag.

[7] Browne, W. J., Goldstein, H. and Rasbash, J. (2001) Multiple membership multiple classification (MMMC) models. Statistical Modelling, 1, 103124.

[8] Browne, W. J., McCleery, R. H. Sheldon, B. C. and Pettifor, R. A. (2007). Using crossclassified multivariate mixed response models with

applications to life history traits in great tits (parus major). Statistical Modelling, 7, 217-238.

[9] Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. The British Journal of Statistical Psychology, 16, 137-163.

[10] Davidov, E., H. Dulmer, E. Schluter and P. Schmidt (2012) Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. Journal of Cross-Cultural Psychology. (in press).

[11] De Boeck, P. (2008). Random item IRT models. Psychometrika, 73, 533559.

[12] De Jong, M.G., Steenkamp, J.-B.E.M. (2009) Finite Mixture Multilevel Multidimensional Ordinal IRT Models for Large Scale Cross-Cultural Research, Psychometrika, 75, 332.

[13] De Jong, M.G., Steenkamp, J.-B.E.M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. Journal of Consumer Research, 34, 260278.

[14] Fielding, A., & Goldstein, H. (2006). Cross-classified and Multiple Membership Structures in Multilevel Models: An Introduction and Review. Research Report RR791. Birmingham, UK: Department for Education and Skills, University of Birmingham.

[15] Goldstein, H. (2011). Multilevel Statistical Models, 4rd Edition. London, Arnold.

[16] Gonzalez, De Boeck, Tuerlinckx (2008) A Double-Structure Structural Equation Model for Three-Mode Data. Psychological Methods, 337 - 353.

[17] Fox, J.P. (2010) Bayesian Item Response Theory. Springer.

[18] Luo, W., & Kwok, O-M. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. Multivariate Behavioral Research, 44, 182-212.

[19] Luo, W., & Kwok, O.-M. (2012). The Consequences of Ignoring Individuals' Mobility in Multilevel Growth Models: A Monte Carlo Study. Journal of Educational and Behaviorial Statistics, 37, 3156.

[20] Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. American Journal of Community Psychology, 27, 599-642.

[21] Jahng S., Wood, P. K.,& Trull, T. J., (2008) Analysis of Affective Instability in Ecological Momentary Assessment: Indices Using Successive Difference and Group Comparison via Multilevel Modeling. Psychological Methods, 13, 354-375.

[22] Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S.E. Embretson & S.L. Hershberger et al. (Eds.). The new rules of measurement: What every psychologist and educator should know. (pp. 129-152). Mahwah, NJ, USA: Lawrence Erlbaum.

[23] Marsh, H. W. and Hocevar, D. (1991) Multidimensional perspective on students evaluations of teaching effectiveness: the generality of factor structures across academic discipline, instructor level, and course level, Teaching and Teacher Education, 7, 918.

[24] Muthén, B. and Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. Forthcoming in Psychological Methods.

[25] Muthén, B. O. (1994). Multilevel covariance structure analysis. Sociological Methods & Research 22(3), 376-398.

[26] Muthén, L.K. and Muthén, B.O. (1998-2010). Mplus Users Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén

[27] Rasbash, J. and Goldstein, H. (1994) Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. Journal of Behavioural Statistics 19, 337-350.

[28] Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross sectional and longitudinal research. Journal of Educational Statistics 18: 321-349.

[29] Raudenbush, S.W. and Bryk, A.S. (2002). Hierarchical Linear Models. Thousand Oaks: Sage Publications.

[30] Verhagen, A. J. & Fox, J.-P. (2012) Bayesian Tests of Measurement Invariance. Conditionally accepted in British Journal of Mathematical and Statistical Psychology.

[31] Walls, T.A. & Schafer, J.L. (2006). Models for Intensive Longitudinal Data. New York : Oxford University Press.