Multi-Dimensional, Multi-Level, and Multi-Timepoint Item Response Modeling.

Bengt Muthén & Tihomir Asparouhov

In van der Linden, W. J., Handbook of Item Response Theory. Volume One. Models, pp. 527-539. Boca Raton: CRC Press

1 Introduction

Item response modeling in the general latent variable framework of the Mplus program (Muthén & Muthén, 2012) offers many unique features including multidimensional analysis (Asparouhov & Muthén, 2012a); two-level, three-level, and cross-classified analysis (Asparouhov & Muthén, 2012b); mixture modeling (Muthén, 2008; Muthén & Asparouhov, 2009); and multilevel mixture modeling (Asparouhov & Muthén, 2008; Henry & Muthén, 2010). This chapter presents a subset of the Mplus item response modeling techniques through the analysis of an example with three features common in behavioral science applications: multiple latent variable dimensions, multilevel data, and multiple timepoints. The dimensionality of a measurement instrument with categorical items is investigated using exploratory factor analysis with bi-factor rotation. Variation across students and classrooms is investigated using two-level exploratory and confirmatory bi-factor models. Change over grades is investigated using a longitudinal two-level model. The analyses are carried out using weighted least-squares, maximum-likelihood, and Bayesian analysis. The strengths of weighted least-squares and Bayesian estimation as a complement to maximum-likelihood for this high-dimensional application are discussed. Mplus scripts for all analyses are available at www.statmodel.com.

As a motivating example, consider a teacher-rated measurement instrument capturing aggressive-disruptive behavior among a sample of U.S. students in Baltimore public schools (Ialongo et al., 1999). A total of 362 boys was observed in 27 classrooms in the Fall of Grade 1 and Spring of Grade 3. The instrument consisted of 13 items scored as 1 (Almost Never) through 6 (Almost Always).

The items and the percentage in each response category are shown in Table 1 for Fall Grade 1. The item distribution is very skewed with a high percentage in the Almost Never category. The responses are modeled as ordered categorical. It is of interest to study the dimensionality of the instrument, to explore response variation due to both students and classrooms, and to study changes over the grades.

[Table 1 about here.]

2 Modeling

2.1 Single-Level Modeling

Let U_{pi} be the response for person p on an ordered polytomous item i with categories $a = 1, 2, \ldots, A$, and express the item probabilities for this item as functions of D factors θ_{pd} $(d = 1, 2, \ldots, D)$ as follows,

$$P(U_{pi} = a|\theta_{p1}, \theta_{p2}, \ldots, \theta_{pD}) = F[\tau_{ia} - \sum_{d=1}^{D} \lambda_{id}\, \theta_{pd}] - F[\tau_{ia-1} - \sum_{d=1}^{D} \lambda_{id}\, \theta_{pd}], \quad (1)$$

where F is either the logistic or standard normal distribution function, corresponding to logistic and probit regression. In statistics, this is referred to as a proportional odds model (Agresti, 2002), whereas in psychometrics it is referred to as a graded response model. When $a = 0$, the threshold parameter $\tau_{i0} = -\infty$, resulting in $F = 0$, and when $a = A$, $\tau_{iA} = \infty$, resulting in $F = 1$. As usual, conditional independence is assumed among the items given the factors and the factors are assumed to have a multivariate normal distribution.

It is also useful to view the model as an equivalent latent response variable model, motivating the use of a threshold formulation. Consider I continuous latent response variables U^*_{pi} $(i = 1, 2, \ldots, I)$ for person p, following a linear factor model with D factors

$$U^*_{pi} = \nu_i + \sum_{d=1}^{D} \lambda_{id}\, \theta_{pd} + \epsilon_{pi}, \tag{2}$$

with a threshold formulation such as for a three-category ordered categorical item,

$$U_{pi} = \begin{cases} 0 & \text{if } U^*_{pi} \leq \tau_{i1} \\ 1 & \text{if } \tau_{i1} < U^*_{pi} \leq \tau_{i2} \\ 2 & \text{if } U^*_{pi} > \tau_{i2} \end{cases}$$

The intercept parameters ν are typically fixed at zero given that they cannot be separately identified from the thresholds τ. With normal factors θ and residuals ϵ the U^* variables have a multivariate normal distribution where the association among the items can be described via latent response variable (LRV) polychoric correlations (see, e.g., Muthén, 1978, 1984). This corresponds to a probit or normal ogive IRT model. With logistic density for the residuals, a logit IRT model is obtained.

2.2 Two-Level Modeling

The two-level model is conveniently expressed in terms of the continuous latent response variables U^*. For person p, item i, and cluster (classroom) j, consider the two-level, random measurement parameter IRT model (see, e.g., Fox, 2010;

Fox & Glas, vol. 1, chap.24)

$$U^*_{pij} = \nu_{ij} + \sum_{d=1}^{D} \lambda_{ijd}\, \theta_{pjd} + \epsilon_{pij}, \tag{3}$$

$$\nu_{ij} = \nu_i + \delta_{\nu ij}, \tag{4}$$

$$\lambda_{ijd} = \lambda_{id} + \delta_{\lambda ijd}, \tag{5}$$

$$\theta_{pjd} = \eta_{jd} + \zeta_{pjd}, \tag{6}$$

so that the factors have between- and within-cluster variation. Given that all loadings are free, one may set the within-cluster variance at one for identification purposes, $\mathrm{V}(\zeta_{pjd}) = 1$. This model can be estimated in Mplus using Bayes as discussed in Asparouhov & Muthén (2012b). When the measurement parameters of ν (or τ) and λ are not varying across clusters but are fixed parameters, the variances of the δ residuals are zero.

It may be noted in the above IRT model that the same loading multiplies both the within- and between-cluster parts of each factor. An alternative model has been put forward in the factor analysis tradition where this restriction is relaxed, acknowledging that some applications call for not only the loadings but also the number of factors to be different on the two levels (see, e.g. Cronbach, 1976; Harnqvist, 1978; Harnqvist et al., 1994). For the case of fixed loading parameters this model may be expressed as

$$U^*_{pij} = \nu_{ij} + \sum_{d=1}^{D_W} \lambda_{Wid}\, \theta_{Wpd} + \epsilon_{Wpij}, \tag{7}$$

$$\nu_{ij} = \nu_i + \sum_{d=1}^{D_B} \lambda_{Bid}\, \theta_{Bjd} + \epsilon_{Bij}, \tag{8}$$

which expresses the two-level factor model as a random intercepts (threshold) model. In this way, there is a separate factor analysis structure on each level. This type of modeling follows the tradition of two-level factor analysis for continuous responses; see, e.g. Goldstein and McDonald (1988), McDonald and Goldstein (1989), and Longford and Muthén (1992). For a discussion of two-level covariance structure modeling in Mplus, see Muthén (1994). For the case of $D_W = D_B$ and $\lambda_{Wid} = \lambda_{Bid}$, the model of (7) - (8) is the same as the model of (3) - (6) with zero variance for $\delta_{\lambda ijd}$, that is, with fixed as opposed to random loadings.

3 Estimation

Three estimators are considered in this chapter, maximum-likelihood, weighted least-squares, and Bayes. Maximum-likelihood (ML) estimation of this model is well-known (see, e.g., Baker and Kim, 2004) and is not described here. Weighted least-squares and Bayes are described only very briefly, pointing to available background references.

3.1 Weighted Least Squares (WLSMV)

Let $\boldsymbol{s}$ define a vector of all the thresholds and LRV correlations estimated from the sample, let $\boldsymbol{\sigma}$ refer to the corresponding population correlations expressed in terms of the model parameters, and let $\boldsymbol{W}$ denote the large-sample covariance matrix for $\boldsymbol{s}$. The weighted least-squares estimator minimizes the sums of squares of the differences between $\boldsymbol{s}$ and $\boldsymbol{\sigma}$, where differences with larger variance are given less weight. This is accomplished by minimizing the following fitting function with respect to the model parameters (Muthén, 1978; Muthén, 1984; Muthén &

Satorra, 1995; Muthén et al., 1997)

$$F = (\boldsymbol{s} - \boldsymbol{\sigma})' \, diag(\boldsymbol{W})^{-1} \, (\boldsymbol{s} - \boldsymbol{\sigma}). \tag{9}$$

Here, $diag(\boldsymbol{W})$ denotes the diagonal of the weight matrix, that is, using only the variances. The full weight matrix is, however, used for χ^2 testing of model fit and for standard error calculations (Muthén et al., 1997; Asparouhov & Muthén, 2010a). This estimator is referred to as WLSMV. Modification indices in line with those typically used for continuous items are available also with weighted least squares (Asparouhov & Muthén, 2010a) and are useful for finding evidence of model misfit such as correlated residuals and non-invariance across groups.

Two-level analysis of categorical data can be carried out by the two-level weighted least-squares (WLSMV) estimator in Mplus developed by Asparouhov and Muthén (2007). This uses the model version of (7) - (8). Two-level factor analysis may involve many latent variables and ML is therefore cumbersome due to many dimensions of numerical integration. Bayesian estimation is feasible, but two-level WLSMV is a simple and much faster procedure suitable for initial analysis. The computational demand is virtually independent of the number of latent variables because high-dimensional integration is replaced by multiple instances of one- and two-dimensional integration using a second-order information approach of WLSMV in line with the Muthen (1984) single-level WLSMV. This implies that residuals can be correlated and that model fit to the LRV structure can be obtained by chi-square testing.

3.2 Bayesian Estimation

As applied to the item response setting, the Bayes implementation in Mplus considers multivariate normal latent response variables as in (2); see, e.g., Johnson and Albert (1999) or Fox (2010). This relies on probit relations where the residual variances are fixed at one as in the maximum-likelihood parameterization using probit. In addition to parameters, latent variables, and missing data, the Bayes iterations consider the latent response variables as unknown to obtain a well-performing Markov Chain Monte Carlo approach. Posterior distributions for these latent response variables are obtained as a side product. Posterior predictive checking is available for the LRV structure (Asparouhov & Muthén, 2010b). The default for Bayesian estimation in Mplus is to use non-informative priors, but informative priors are easily specified. Like MLE, Bayesian estimation is a full-information approach and with non-informative priors Bayes gives asymptotically the same estimates as MLE. Bayesian estimation of latent variable models with categorical items as in IRT is, however, advantageous to MLE due to numerical integration required for the latter, which is slow or prohibitive with many dimensions of integration due to many latent variables. From a practical point of view, Bayesian analysis with non-informative priors can in such cases be seen as an approach to getting estimates close to those of MLE if ML estimates could have been computed. Bayesian estimation is particularly useful for two-level item response models due to many latent variable dimensions. For technical aspects, see Asparouhov and Muthén (2010b).

4 Empirical Examples

Returning to our motivating example, this section goes through a series of analyses to demonstrate the flexibility of the item response modeling in Mplus. For the software implementations of these analyses, see Muthén and Muthén (vol. 3, chap. 29).

4.1 Item Bi-Factor Exploratory Factor Analysis

Using ML, WLSMV, and Bayes estimation, Muthén et al. (2012) found that a three-factor exploratory factor analysis (EFA) model is suitable for the aggressive-disruptive items described in Section 1. The three factors correspond to verbally-oriented, property-oriented, and person-oriented aggressive-disruptive behavior. An alternative EFA model with identical fit to the data is a bi-factor model that has both a general factor influencing all items and also specific factors, uncorrelated with the general factor, which influence sets of items. Drawing on Jennrich and Bentler (2011, 2012), it is possible to carry out a bi-factor EFA using an extension to categorical items implemented in Mplus. Assuming unidimensionality, a summed score of the 13 items of this measurement instrument has previously been used (Ialongo et al., 1999). The general factor of the bi-factor model is related to such an overall score, but it is of interest to also study additional specific dimensions.

The bi-factor EFA estimates in Table 2 pertain to the Fall Grade 1 data and are obtained using the WLSMV estimator with a bi-factor Geomin rotation allowing correlated specific factors. The table shows that all items load significantly on the general factor with approximately equal loadings. The first specific factor,

labeled number 2, has significant and positive loadings for the items stubborn and loses temper, which may be weakly indicative of a verbally-oriented aggressive-disruptive specific factor. There are, however, several large negative loadings which make the interpretation less clear. The second specific factor, labeled number 3, has significant and positive loadings for the items harms others, fights, teases classmates, and fights with classmates, which may be indicative of a person-oriented aggressive-disruptive specific factor. The two specific factors have a small significant positive correlation. Modification indices suggest a few correlated residuals, but including them does not alter the original estimates in any important way.

[Table 2 about here.]

4.2 Two-Level Item Bi-Factor Exploratory Factor Analysis

The data described in Section 1 are obtained from students in 27 classrooms. This multilevel structure was ignored in the previous analysis. It is, however, possible to carry out bi-factor EFA also for two-level categorical data using two-level weighted least-squares (WLSMV) estimation in Mplus developed by Asparouhov and Muthén (2007). This uses the model version of (7) - (8).

Table 3 shows the two-level WLSMV solution for the student-level part of the model using a model with two specific factors on each level. The two-level pattern is much clearer than in the single-level analysis ignoring clustering shown in Table 2. The large negative loadings for the specific factors have largely disappeared. The first specific factor, labeled 2, now has more significant

positive loadings and thereby more clearly defines the factor as a verbally-oriented aggressive-disruptive specific factor. The second specific factor, labeled number 3, is largely unchanged compared to the single-level analysis representing a person-oriented aggressive-disruptive specific factor.

The classroom-level loadings do not give an interpretable picture. Two of the three factors do not have any significant loadings so that it is not clear that the loadings are invariant across the two levels or that three factors are needed on the classroom level.

[Table 3 about here.]

4.3 Two-Level Item Bi-Factor Confirmatory Factor Analysis

In this section, the analysis is changed from exploratory factor analysis to confirmatory factor analysis (CFA). Using the factor pattern shown in Table 3, a CFA bi-factor model is specified where the specific factors corresponding to verbal- and person-oriented aggressive-disruptive behavior are specified to be measured by only the factor loadings with asterisks.

Using two-level WLSMV estimation, a model is considered with the same loading pattern on the student and classroom levels, but not restricting the loadings to be equal across the two levels. Using this model, Wald testing of loading equality across levels is easily carried out in Mplus and it is found that equality cannot be rejected.

As an initial step in a series of further analyses, Bayesian estimation of the two-level bi-factor CFA model with loading invariance across levels is carried out

and the estimates shown in Table 4 (the loadings are in a different metric than in Table 3 where the U^* variances are set to one). Because of the loading invariance across levels, it is possible to study the decomposition of factor variance for the two levels. The percentage due to the classroom variance may be seen as an indicator of heterogeneity of aggressive-disruptive classroom environment. It is seen that 24% ($100 \times 0.322/(1 + 0.322)$ of the general factor variance is due to variation across classrooms, with 27% for the verbal factor and 35% for the person factor.

[Table 4 about here.]

4.4 Two-Level Item Bi-Factor Confirmatory Factor Analysis with Random Factor Loadings

This section changes the two-level model from the model version of (7) - (8) to the model version of (3) - (6). Because equality of factor loadings across the two levels is imposed in line with the previous section, the key difference between the two model types is that the factor loadings are now allowed to vary across the clusters, in this case the classrooms. Technical aspects of this model are described in Asparouhov and Muthén (2012b) using several model variations. The specification makes an attempt to absorb as much of the factor loading variation as possible in factor variance differences across clusters.

The loadings are found to have substantial variation across the 27 classrooms. In other respects, however, the results are close to those of Table 4. The average loadings are similar for the general factor and the factor variances on the classroom level are similar: 0.348 for the general factor, 0.364 for the verbal factor, and

0.374 for the person factor. Given that the general conclusions are not altered, subsequent modeling holds loadings equal across clusters.

4.5 Longitudinal Two-Level Item Bi-Factor Confirmatory Factor Analysis

The 13 aggressive-disruptive behavior items were measured not only in Grade 1 but also in the two subsequent grades. This section discusses longitudinal item response modeling with a focus on changes across time in factor means and variances. Joint analyses of Grade 1 and Grade 3 are carried out, while at the same time taking into account the classroom clustering. In this sense, three-level data are considered. The analyses presented here, however, will be carried out as two-level modeling because a wide format approach is taken for the longitudinal part of the model, formulating a model for the multivariate vector of 2×13 items.

As is typical in longitudinal studies, many students measured in Grade 1 are missing in Grade 3. In these data 28% of the students are missing. In such cases it is important to be able to draw on the missing data assumption of MAR (Little & Rubin, 2002), requiring the full-information estimation approaches of ML or Bayes. MCAR cannot be taken for granted, which is assumed by the WLSMV estimator due to using information from only pairs of variables. WLSMV is, however, without such a limitation when used together with a first step of multiple imputation of the missing data (for multiple imputation using Mplus, see Asparouhov & Muthén 2010c).

In the longitudinal setting, this application requires a special type of multilevel modeling of the classroom clustering. Classroom membership pertains to Grade

1 classrooms, while the students are spread over many different classrooms by Grade 3. An indication of the effect of Grade 1 classroom clustering on Grade 1 and Grade 3 outcomes is obtained using intraclass correlations for the U^* variables behind the 2×13 items. Table 5 presents the intraclass correlations using the two-level WLSMV and Bayes estimators. This information is intractable to obtain with ML due to requiring an unrestricted model for the U^* variables on both the student and classroom level. Because a probit response function is used, these intraclass correlations are computed with a unit within-cluster variance as

$$icc = \sigma_B^2/(1 + \sigma_B^2), \tag{10}$$

where σ_B^2 is the between-level variance of the random intercept for the item.

[Table 5 about here.]

Table 5 shows that the Grade 1 items have sizeable intraclass correlations, but that by Grade 3 the Grade 1 classroom effect has largely disappeared. Grade 3 classroom clustering is presumably still present because many students are in the same classroom, but the overall clustering effect in Grade 3 is probably smaller due to fewer students of this cohort being in the same classroom in Grade 3. Such a Grade 3 clustering effect is ignored here.

The longitudinal model to be used draws on the Grade 1 model of Section 4.3 using a two-level item bi-factor confirmatory model with equal loadings across the two levels. The model is extended to include Grade 3 responses as follows. For the student level the same bi-factor model is specified, holding loadings equal to those in Grade 1. For the classroom level, the Grade 3 items are influenced by the Grade

1 general factor with no classroom-level factors added for Grade 3. This is in line with the small intraclass correlations for Grade 3, where the expectation is that the classroom-level loadings for the Grade 3 items on the general factor of Grade 1 will be small. In this way, the longitudinal model has three student-level factors in Grade 1, three student-level factors in Grade 3, and three classroom-level factors in Grade 1 for a total of nine factors. In addition, classroom-level residual variances for the items in Grade 1 and Grade 3 add further latent variable dimensions for a total of 35. The student-level factors are allowed to correlate across grades.

The longitudinal model is also extended to consider changes across time in the means of both the general and the specific factors. This is accomplished by also imposing measurement invariance for the item threshold across Grade 1 and Grade 3. Factor means are fixed at zero for Grade 1 and estimated for Grade 3.

For Bayes estimation, the fact that the model is high-dimensional does not present a problem. For ML estimation, however, this leads to intractable computations. Reducing the dimensions of the model to nine by fixing the classroom-level residual variances at zero, it is possible to carry out the ML computations using Monte Carlo integration (see Asparouhov & Muthén, 2012a) with a total of 100,000 integration points divided into 2,174 points for the six student-level dimensions and 46 points for the three classroom-level dimensions. This computationally-heavy ML analysis is, however, 10 times slower than the Bayes analysis of the full model.

The results of the Bayesian analysis are presented in Table 6. For simplicity, the items are dichotomized in this analysis. The items are dichotomized between the two most frequent item categories of Almost Never and Rarely (see Table 1).

Table 6 shows that the student-level factor variances decrease from Grade 1

to Grade 3 so that student behavior becomes more homogeneous. The means of the general and verbal factors increase significantly from Grade 1 to Grade 3, while the increase for the person factor is not significant. In terms of Grade 1 factor standard deviations, the increase in factor mean from Grade 1 to Grade 3 is 0.298 for the general factor and 0.563 for the verbal factor, indicating that the increase in aggressive-disruptive behavior is mostly due to increased verbally-oriented aggressive-disruptive behavior.

These analyses may form the basis for a multiple-indicator growth model across several grades where growth is considered for both the general and the specific factors. Previous growth analyses for these data have been carried out on the sum of the items assuming unidimensionality. Using growth mixture modeling, these analyses have uncovered different latent classes of trajectories for which an intervention has different effects (Muthén et al., 2002; Muthén & Asparouhov, 2009). Such a finite mixture generalization is also possible with the multidimensional, multilevel item response modeling considered here.

[Table 6 about here.]

5 Conclusions

The analysis of the aggressive-disruptive behavior application exemplifies the flexibility of item response modeling in the Mplus framework. High-dimensional exploratory, confirmatory, multilevel, and longitudinal analyses are possible using a combination of weighted least-squares, maximum-likelihood, and Bayesian estimation.

Due to lack of space, many more analysis possibilities relevant to this

application are excluded from the discussion. Exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) can be used for multiple-group EFA of male and female students with varying degrees of invariance of measurement and structural parameters. ESEM can also be used in a longitudinal analysis to study measurement invariance across grades. Bayesian EFA and two-tier modeling can be carried out as discussed in Asparouhov and Muthén (2012a). Bayesian structural equation modeling (BSEM; Muthén & Asparouhov, 2012) can be used to allow cross-loadings in the confirmatory analysis, using informative zero-mean, small-variance priors for parameters that are not identified in maximum-likelihood analysis. Gender differences can be studied in two-level analysis that allows within-cluster groupings as discussed in Asparouhov and Muthén (2012c). Mplus applications of more general latent variable models with random subjects, random items, random contexts, and random parameters are discussed in Asparouhov and Muthén (2012b).

References

Agresti, A. (2002). *Categorical data analysis.* Second edition. New York: John Wiley & Sons.

Asparouhov, T., & Muthén, B. (2007, July). *Computationally efficient estimation of multilevel high-dimensional latent variable models.* Paper presented at the Joint Statistical Meetings, Salt Lake City, Utah.

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing, Inc.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397-438.

Asparouhov, T., & Muthén, B. (2010a). *Simple second order chi-square correction.* Unpublished manuscript.

Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis using Mplus: Technical implementation.* Unpublished manuscript.

Asparouhov, T., & Muthén, B. (2010c). *Multiple imputation with Mplus.* Unpublished manuscript.

Asparouhov, T., & Muthén, B. (2012a). *Comparison of computational methods for high-dimensional item factor analysis.* Manuscript submitted for publication.

Asparouhov, T., & Muthén, B. (2012b). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Unpublished manuscript.

Asparouhov, T., & Muthén, B. (2012c). *Multiple group multilevel analysis. Mplus Web Notes: No. 16. November 15, 2012.* Manuscript submitted for

publication.

Baker, F. B., & Kim, S. H. (2004). *Item response theory. Parameter estimation techniques.* New York: Marcel Dekker.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis.* Unpublished manuscript.

Fox, J.P. (2010). *Bayesian item response modeling.* New York: Springer.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455-467.

Harnqvist, K. (1978). Primary mental abilities of collective and individual levels. *Journal of Educational Psychology, 70,* 706-716.

Harnqvist, K., Gustafsson, J. E., Muthén, B., & Nelson, G. (1994). Hierarchical models of ability at class and individual levels. *Intelligence, 18,* 165-187.

Henry, K., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling, 17,* 193-215.

Ialongo, L. N., Werthamer, S., Kellam, S. K., 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,* 599-641.

Jennrich, R. I. & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76,* 537- 549.

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika, 77,* 442-454.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling.* New York: Springer.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data.* Second edition. New York: John Wiley and Sons.

Longford, N. T., & Muthén, B. (1992). Factor analysis for clustered observations. *Psychometrika, 57,* 581-597.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42,* 215-232.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551-560.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, B. (1994). Multilevel covariance structure analysis. In J. Hox & I. Kreft (Eds.), *Multilevel Modeling, a special issue of Sociological Methods & Research, 22,* 376-398.

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age Publishing, Inc.

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

Muthén, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods, 17,* 313-335.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., Wang, C.

P., Kellam, S., Carlin, J., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3,* 459-475.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide.* Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, B., Muthén, L., & Asparouhov, T. (2012). *Regression analysis, factor analysis, and structural equation modeling using Mplus.* Book in preparation.

Muthén, B., & Satorra, A. (1995). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60,* 489-503.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

List of Tables

Table 1: Response percentages for aggression items of $n = 363$ cohort 3 males in Fall of Grade 1

	Almost Never (scored as 1)	Rarely (scored as 2)	Sometimes (scored as 3)	Often (scored as 4)	Very Often (scored as 5)	Almost Always (scored as 6)
stubborn	42.5	21.3	18.5	7.2	6.4	4.1
breaks rules	37.6	16.0	22.7	7.5	8.3	8.0
harms others and property	69.3	12.4	9.40	3.9	2.5	2.5
breaks things	79.8	6.60	5.20	3.9	3.6	0.8
yells at others	61.9	14.1	11.9	5.8	4.1	2.2
takes others' property	72.9	9.70	10.8	2.5	2.2	1.9
fights	60.5	13.8	13.5	5.5	3.0	3.6
harms property	74.9	9.90	9.10	2.8	2.8	0.6
lies	72.4	12.4	8.00	2.8	3.3	1.1
talks back to adults	79.6	9.70	7.80	1.4	0.8	1.4
teases classmates	55.0	14.4	17.7	7.2	4.4	1.4
fights with classmates	67.4	12.4	10.2	5.0	3.3	1.7
loses temper	61.6	15.5	13.8	4.7	3.0	1.4

Table 2: Bi-factor EFA solution using WLSMV (asterisks indicate significance at the 5% level)

	1	2	3
Bi-Geomin rotated loadings			
stubborn	0.718*	**0.398***	0.013
breaks rules	0.796*	0.099	0.107
harms others and property	0.827*	-0.197*	**0.198***
breaks things	0.890*	-0.330*	0.007
yells at others	0.842*	0.180	-0.013
takes others' property	0.848*	-0.242	-0.017
fights	0.892*	-0.040	**0.367***
harms property	0.921*	-0.289	-0.020
lies	0.906*	-0.049	-0.128*
talks back to adults	0.870*	0.255	-0.116
teases classmates	0.806*	0.008	**0.178***
fights with classmates	0.883*	0.060	**0.399***
loses temper	0.826*	**0.273***	0.003
	1	2	3
Bi-Geomin factor correlations			
1	1.000		
2	0.000	1.000	
3	0.000	0.115*	1.000

Table 3: Two-level analysis using bi-factor EFA and the WLSMV estimator. Student-level results.

	1	2	3
Geomin rotated loadings			
stubborn	0.699*	**0.360***	-0.011
breaks rules	0.829*	0.079	0.054
harms others and property	0.876*	-0.053	-0.021
breaks things	0.918*	-0.025	-0.211*
yells at others	0.795*	**0.293***	0.009
takes others' property	0.875*	-0.134*	-0.015
fights	0.927*	-0.043	**0.287***
harms property	0.944*	-0.003	-0.125*
lies	0.894*	0.066	-0.070
talks back to adults	0.837*	**0.349***	-0.004
teases classmates	0.814*	0.033	**0.187***
fights with classmates	0.919*	0.009	**0.314***
loses temper	0.780*	**0.390***	0.009
	1	2	3
Geomin factor correlations			
1	1.000		
2	0.000	1.000	
3	0.000	0.263*	1.000

Table 4: Two-level analysis using bi-factor CFA and the Bayes estimator.

	G	Verbal	Person
Factor loadings			
stubborn	1.090*	0.633*	0.000
breaks rules	1.519*	0.000	0.000
harms others and property	1.839*	0.000	0.000
breaks things	2.699*	0.000	0.000
yells at others	1.543*	0.668*	0.000
takes others' property	1.915*	0.000	0.000
fights	3.525*	0.000	1.512*
harms property	3.452*	0.000	0.000
lies	2.166*	0.000	0.000
talks back to adults	2.000*	0.884*	0.000
teases classmates	1.511*	0.000	0.436*
fights with classmates	4.534*	0.000	2.253*
loses temper	1.689*	1.084*	0.000
	G	Verbal	Person
Factor variances			
Student-level variances	1.000	1.000	1.000
Classroom-level variances	0.322	0.375	0.547

Table 5: Intraclass correlations for Grade 1 and Grade 3 responses estimated with WLSMV and Bayes

	Grade 1 WLSMV	Grade 1 Bayes	Grade 3 WLSMV	Grade 3 Bayes
stubborn	0.110	0.099	0.000	0.080
breaks rules	0.121	0.105	0.000	0.072
harms others and property	0.208	0.138	0.000	0.075
breaks things	0.380	0.222	0.015	0.104
yells at others	0.215	0.142	0.000	0.070
takes others' property	0.252	0.179	0.000	0.074
fights	0.159	0.100	0.000	0.072
harms property	0.314	0.202	0.001	0.083
lies	0.211	0.172	0.000	0.070
talks back to adults	0.143	0.122	0.000	0.068
teases classmates	0.177	0.126	0.026	0.089
fights with classmates	0.160	0.100	0.000	0.073
loses temper	0.171	0.119	0.000	0.078

Table 6: Grade 1 - Grade 3 longitudinal two-level analysis using bi-factor CFA and the Bayes estimator. Dichotomized items

	G	Verbal	Person
Factor loadings			
stubborn	1.270*	0.462*	0.000
breaks rules	1.783*	0.000	0.000
harms others and property	1.857*	0.000	0.000
breaks things	1.838*	0.000	0.000
yells at others	1.836*	0.327	0.000
takes others' property	2.295*	0.000	0.000
fights	3.106*	0.000	1.110*
harms property	2.758*	0.000	0.000
lies	2.815*	0.000	0.000
talks back to adults	2.684*	1.571*	0.000
teases classmates	1.649*	0.000	0.442*
fights with classmates	3.397*	0.000	1.318*
loses temper	1.708*	0.610*	0.000
	G	Verbal	Person
Factor variances			
Student-level Grade 1	1.000	1.000	1.000
Student-level Grade 3	0.895	0.427	0.620
Classroom-level Grade 1	0.203	0.318	0.418
Factor means			
Grade 1	0.000	0.000	0.000
Grade 3	0.327*	0.646*	0.286
Factor correlation			
G for Grade 1 with Grade 3	0.349*		