

The Efficacy of Common Fit Indices for Enumerating Classes in Growth Mixture Models When Nested Data Structure Is Ignored: A Monte Carlo Study

SAGE Open
January-March 2017: 1–19
© The Author(s) 2017
DOI: 10.1177/2158244017700459
journals.sagepub.com/home/sgo


Qi Chen¹, Wen Luo², Gregory J. Palardy³, Ryan Glaman¹,
and Amber McEnturff¹

Abstract

Growth mixture model (GMM) is a flexible statistical technique for analyzing longitudinal data when there are unknown heterogeneous subpopulations with different growth trajectories. When individuals are nested within clusters, multilevel growth mixture model (MGMM) should be used to account for the clustering effect. A review of recent literature shows that a higher level of nesting was described in 43% of articles using GMM, none of which used MGMM to account for the clustered data. We conjecture that researchers sometimes ignore the higher level to reduce analytical complexity, but in other situations, ignoring the nesting is unavoidable. This Monte Carlo study investigated whether the correct number of classes can still be retrieved when a higher level of nesting in MGMM is ignored. We investigated six commonly used model selection indices: Akaike information criterion (AIC), consistent AIC (CAIC), Bayesian information criterion (BIC), sample size-adjusted BIC (SABIC), Vuong–Lo–Mendell–Rubin likelihood ratio test (VLMR), and adjusted Lo–Mendell–Rubin likelihood ratio test (ALMR). Results showed that accuracy of class enumeration decreased for all six indices when the higher level is ignored. BIC, CAIC, and SABIC were the most effective model selection indices under the misspecified model. BIC and CAIC were preferable when sample size was large and/or intraclass correlation (ICC) was small, whereas SABIC performed better when sample size was small and/or ICC was large. In addition, SABIC and VLMR/ALMR tended to overextract the number of classes when there are more than two subpopulations and the sample size is large.

Keywords

growth mixture model, multilevel model, class enumeration, fit index, clustered data

Introduction

Multilevel growth mixture model (MGMM) is a relatively new modeling technique for extracting unknown subpopulations in multilevel longitudinal data. This technique integrates multilevel modeling, finite mixture modeling, and structural equation modeling (Asparouhov & Muthén, 2008; B. Muthén, 2004). The multilevel aspect of MGMM is attractive to applied researchers because longitudinal data are often collected through cluster sampling, which creates multilevel data structure with repeated measures nested within individuals and individuals further nested within organizations. Some examples are students nested within classrooms/schools/neighborhoods (e.g., Dettmers, Trautwein, Lüdtke, Kunter, & Baumert, 2010), children/couples nested within families (e.g., Jenkins, Dunn, O'Connor, Rasbash, & Behnke, 2005; Pruchno, Wilson-Genderson, & Cartwright, 2009), individuals nested within countries (e.g., Matsumoto, Nezlek, & Koopmann, 2007), clients nested within therapists

(e.g., Marcus, Kashy, & Baldwin, 2009), and individuals nested within organizations (e.g., Vancouver, 1997). With multilevel longitudinal data, unknown subpopulations can be extracted at the individual level as well as the organization level (Palardy & Vermunt, 2010). In addition, researchers can study the associations between organizational characteristics and individual growth patterns (note: in this article, the terms class and subpopulation are used interchangeably). It should be noted that other methods, such as item response theory (e.g., Bartolucci, Pennoni, & Vittadini, 2011), can be used to account for longitudinal data structure in the analysis

¹University of North Texas, Denton, USA

²Texas A&M University, College Station, USA

³University of California, Riverside, USA

Corresponding Author:

Qi Chen, Department of Educational Psychology, University of North Texas, 1155 Union Circle, #311335, Denton, TX 76203-5017, USA.
Email: qi.chen@unt.edu



of latent traits as well, but the focus of the current article is on MGMM. Due to space constraints, interested readers may find more detailed technical information on MGMM in Chen, Kwok, Luo, and Willson (2010). Also note that the current article uses many acronyms throughout; therefore, for reference, a list of these acronyms and what they denote is provided in the appendix.

To investigate the prevalence of the higher level nesting in growth mixture model (GMM), we reviewed 158 substantive articles with 196 GMMs found in the PsycInfo database between 2011 and 2013. Of these, authors described nesting structure in 85 GMMs (43%), none of which used MGMM to address it. Nevertheless, nine articles addressed the clustering by reporting a low degree of clustering or using adjusted standard errors.

Our review shows that multilevel longitudinal data are common, but when applying GMMs, many empirical researchers ignore the highest level of nesting, potentially violating the assumption of independence (e.g., Boscardin, Muthén, Francis, & Baker, 2008; D'Angiulli, Siegel, & Maggi, 2004). Some reasons for ignoring a level of nesting are avoidable, such as reducing analytic complexity and reducing the difficulty in achieving convergence in model estimation, while others are inevitable such as lack of identifiers (IDs) for higher level units.

The literature has demonstrated that if the nonindependence is not accounted for, parameter estimates and standard errors in a multilevel regression model could be biased (Moerbeek, 2004). In MGMM, ignoring a higher level of nesting structure could result in lower classification accuracy, overestimated lower level variance components, and biased standard errors, which affect significance tests for fixed effects (Chen et al., 2010).

In mixture modeling, one important yet challenging issue is extraction of the correct number of latent classes (e.g., Bartolucci & Murphy, 2015; McLachlan & Peel, 2000). In recent years, many researchers have investigated the performance of various model selection indices in identifying the correct number of classes in different types of mixture models with different data structures (i.e., Allua, 2007; Clark & Muthén, 2007; Henson, Reise, & Kim, 2007; Nylund, Asparouhov, & Muthén, 2007; Peugh & Fan, 2012; Tofighi & Enders, 2008; Yang, 2006). Although the types of mixture models used in simulation studies vary and the best performing indices differ, Bayesian information criterion (BIC; Schwartz, 1978), sample size-adjusted BIC (SABIC; Sclove, 1987), and the Vuong-Lo-Mendell-Rubin likelihood ratio test (VLMR; Lo, Mendell, & Rubin, 2001) consistently performed well for model selection with single-level data.

However, the influence of ignoring a level of nesting on class enumeration for MGMM has not yet been fully investigated. Because ignoring a higher level may be inevitable in some circumstances as described above, an understanding of fit index performance in those situations is warranted. As shown in previous studies on multilevel analysis (e.g., Chen et al., 2010; Meyers & Beretvas, 2006; Moerbeek, 2004),

ignoring the highest level data structure results in the redistribution of variance from the ignored level (i.e., the organization/school level) to the adjacent level (i.e., the individual/student level). It is unclear if this redistribution of variance will affect model selection index performance. It is important to determine whether the recommended model selection indices can extract the correct number of classes when ignoring a higher level of nesting structure is inevitable and to provide researchers with recommendations on using these indices.

Purpose of the Study

The purpose of this study is to (a) investigate whether the correct number of classes can be identified using various commonly used model selection indices when the MGMM is misspecified to omit the higher level and (b) identify factors that affect the index performance for class enumeration when the model is misspecified. The current study extends work by Chen et al. (2010) in several ways. First, whereas Chen et al. focused on the accuracy of classification of individuals and the statistical properties of the parameter estimates (i.e., Type I error rate and statistical power) for each subpopulation, *conditional upon* the correct number of classes being identified, the current study addresses whether the correct number of classes can be enumerated using various model selection indices under both true and misspecified models. This is an important advancement for applying the MGMM given that individual class solutions and corresponding class models can only be further examined and interpreted *after* the correct number of classes is identified. To that end, the current study examines the efficacy of six commonly used indices for class enumeration in GMM.

Compared with Chen et al. (2010), the current study has been expanded to examine one additional design factor by including a true model with three latent classes. Furthermore, we apply some suggested cutoff values for comparing different information criteria between competing models (e.g., delta-BIC, Raftery, 1996) whereas previous studies examining the sensitivity of the information criteria did not account for the magnitude of that difference.

We begin by reviewing the development and specification of MGMM, followed by a brief review of various commonly used model selection indices, and a review of recent studies examining the model index performance. In the simulation study, we first examined a two-subpopulation case (Study 1), followed by a three-subpopulation case (Study 2), because the number of subpopulations may affect the model selection indices' performance. In addition to the number of subpopulations, we investigated the effect of another type of model complexity in Study 3.

Brief Review of MGMMs

The development of MGMMs drew upon several lines of research (Palardy & Vermunt, 2010), namely, latent growth

curve modeling (LGCM; Bollen & Curran, 2006), latent class growth analysis (LCGA; Nagin, 1999), and GMM (B. Muthén, 2004). Combining LGCM and LCGA, GMM is a more general modeling framework capable of examining both the unknown heterogeneous subpopulations and the random variation of the latent growth factors within classes. However, GMM does not consider the situation of multi-level data in which individuals are nested within organizations. Hence, GMM cannot handle nonindependence of individuals due to clustering. Existing research has explored methods of accounting for nonindependence of observations in mixture modeling and GMM (Asparouhov, & Muthén, 2008; Ng & McLachlan, 2014; Ng, McLachlan, Wang, Jones, & Ng, 2006). An extension to GMM, MGMM considers nonindependence of individuals by specifying a model for each level of the multilevel data. The model for the individual and organizational levels can be different, depending on whether heterogeneity is assumed and/or random effects at both the individual level and the organizational level growth trajectories are modeled. This article focuses on the more common MGMM with classification at the individual level (e.g., students being classified into different subgroups within schools; patients being classified into different subtypes within clinics) and no classification at the organizational level.

Brief Review of Model Selection Indices

A combination of substantive knowledge and statistical criteria has been recommended for selecting the optimal number of classes in GMM (B. Muthén, 2003). Generally, model selection statistics can be grouped into four categories: (a) information-based criteria, (b) nested model likelihood ratio tests (LRTs), (c) goodness-of-fit measures, and (d) classification-based statistics (Henson et al., 2007; Tofighi & Enders, 2008; Vermunt & Magidson, 2002). Among these categories, the information-based criteria and the nested model LRTs are the most recommended for determining the number of classes (Henson et al., 2007; Nylund et al., 2007; Tofighi & Enders, 2008). Hence, the model selection indices from these two categories are the focus of this study and are briefly reviewed below.

Information criterion (IC) indices are based on the log-likelihood value of a fitted model and typically penalize model complexity and/or take sample size into account. IC usually takes the form of $-2\log L$ plus a penalty and sample size adjustment, where L is the maximized likelihood. The most commonly used indices include Akaike information criterion (AIC; Akaike, 1987), consistent AIC (CAIC; Bozdogan, 1987), Bayesian information criterion (BIC), and sample size-adjusted BIC (SABIC). For a particular sample and model, the $-2\log L$ is constant. Differences in the penalty terms distinguish the indices and may result in different optimal class enumeration solutions. These fit indices are defined below, respectively:

$$\text{AIC} = -2\log L + 2p, \quad (1a)$$

$$\text{CAIC} = -2\log L + p(\log(N) + 1), \quad (1b)$$

$$\text{BIC} = -2\log L + p\log(N), \quad (1c)$$

$$\text{SABIC} = -2\log L + p\log\left(\frac{N+2}{24}\right), \quad (1d)$$

where p is the number of free parameters in the model and N is the number of subjects. Generally, as a model becomes more complex (i.e., more parameters and larger p), the likelihood increases and $-2\log L$ decreases. IC indices favor models with a relatively higher likelihood value and relatively fewer parameters. Thus, lower IC values indicate a better trade-off between model fit and complexity. For a particular sample and model, the $-2\log L$ is constant. However, differences in the penalty functions (e.g., penalizing model complexity) of different model selection indices result in inconsistent class solutions (i.e., different indices may favor different class solutions). Previous research on mixture model estimation also suggests sample size plays a role in penalty calculations (Leroux, 1992), and that penalties must fulfill two criteria for consistent mixture model estimation: First, as N approaches infinity, penalty / N should become closer to zero. Second, as N approaches infinity, $\log(N)$ / penalty should become closer to zero (Keribin, 2000). Note that the CAIC, BIC, and SABIC fulfill these two conditions.

The IC statistics take both model fit and complexity into consideration. Lower values indicate better trade-off between model fit and complexity. Sometimes, the IC difference between two models is so small that the evidence to support one model over the other becomes very weak. Some guidelines for interpreting the absolute IC difference between two models have been proposed. Petras and Masyn (2010) recommended the “elbow criterion” to determine the optimal number of classes when using IC indices (i.e., AIC, BIC, SABIC). Specifically, they recommended graphing the values of IC indices against the increasing number of classes, and looked for the pronounced angle in the plot where the decrease of IC value dropped. As plotting for all replications was unrealistic, we used cutoff criteria that mimic looking for the “elbow point.”

Besides their statistical differences, the AIC and BIC also have different philosophical contexts (Bauer & Curran, 2003; Burnham & Anderson, 2004; Kuha, 2004; Weakliem, 2004). AIC aims at finding the model that minimizes the Kullback–Leibler (K-L) criterion, selecting an approximate model, and providing better predictions of the population parameters. On the contrary, BIC targets the “true” underlying model with the highest posterior probability. It depends on the purpose of model selection and the nature of reality when deciding which model selection index to use. AIC and BIC were designed for different applications, and both applications can arise in multilevel (growth) mixture modeling.

The nested model LRTs include the VLMR, the adjusted Lo–Mendell–Rubin likelihood ratio test (ALMR; Lo et al., 2001), and the bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000). All these statistics are developed using LRT and test the null hypothesis that the restricted model with $k - 1$ classes fits the data as well as the less restricted model with k classes. The test statistic for a likelihood ratio (LR) test is defined by

$$LR = -2 \log \left(\frac{L_0}{L_u} \right),$$

where L_0 and L_u are the maximized likelihood for the more and less restricted models, respectively (Agresti, 1996). Under the context of mixture model, the LR for $(k - 1)$ -class model versus k -class model is not asymptotically distributed as a chi-square, so the normal chi-square difference test is not applicable. Therefore, Lo et al. (2001) derived an approximate reference distribution for the LR in the mixture context by extending Vuong's (1989) work called the Vuong–Lo–Mendell–Rubin likelihood ratio test (VLMR). Furthermore, Lo et al. (2001) proposed an ad hoc adjustment to VLMR (i.e., ALMR—adjusted Lo–Mendell–Rubin likelihood ratio test), which is defined by

$$LR_{adjusted} = \frac{LR}{1 + [(p - q) \log n]^{-1}},$$

where $p = 3k_1 - 1$ and $q = 3k_0 - 1$ for k_0 -component normal mixture and k_1 -component normal mixture (with k_0 and k_1 both to be known constants and $k_0 < k_1$). A small p value (e.g., $p < .05$) indicates that the $(k - 1)$ -class model should be rejected in favor of the k -class model, while a large p value (e.g., $p \geq .05$) indicates the $k - 1$ and k -class solutions fit the data equally well, and the simpler model (with $k - 1$ classes) is preferable. The testing logic of BLRT is similar to that of VLMR and ALMR. BLRT was not considered in this article because it is not available under the MGMM model. Readers may see Nylund et al. (2007) for more details. It should be noted that the conditions for the ALMR theorem are not generally satisfied in the context of mixture models (Jeffries, 2003); nevertheless, the ALMR has been shown to be effective in recovering the number of underlying components (Nylund et al., 2007).

Review of Studies on Performance of Model Selection Indices

Recently, researchers have studied the performance of these model selection indices in nonclustered data in the context of latent class analysis (LCA; Nylund et al., 2007; Yang, 2006), GMM (Nylund et al., 2007; Peugh & Fan, 2012; Tofighi & Enders, 2008), latent profile analysis (Morgan, Hodge, & Baggett, 2016), latent variable mixture model (Henson et al., 2007), factor mixture model (FMM; Nylund et al., 2007), and latent Markov model (Bacci, Pandolfi, & Pennoni, 2014). A

few studies have searched for the optimal model selection indices for clustered/multilevel data in LCA (Clark & Muthén, 2007; Lukociene & Vermunt, 2010) and FMM (Allua, 2007). Table 1 summarizes these studies, showing the models and indices examined, as well as the best performing or recommended fit indices for class enumeration in each study. As shown in the table, although the types of mixture models vary, they seem to agree on the use of SABIC, BIC, and VLMR/ALMR for model selection in single-level data.

A few studies have searched for the optimal model selection indices for clustered/multilevel data. In LCA, using standard error corrections for clustered data, Clark and Muthén (2007) found that for simple structure data (i.e., latent classes with parallel profiles), none of the studied indices performed well, whereas for complex data structure (i.e., latent classes with crossing profiles), BIC and SABIC performed relatively better. For multilevel FMM, which is a multilevel extension of the factor analysis model for cross-sectional datasets with a hierarchical structure, Allua (2007) found that BIC, SABIC, and ALMR performed better in situations when there was only one class in the population; AIC and ALMR performed better in situations when there were two classes. However, Allua (2007) did not identify any consistently well-performing model selection index for the multilevel FMM. Lukociene and Vermunt (2010) compared the performance of alternate fit indices in multilevel mixture model with focus on determining the true number of mixture components at the organization level. They raised the interesting point that the N for BIC and CAIC computations is unambiguous for single-level mixture evaluation but ambiguous for multilevel mixture evaluation. Their results supported defining N as the number of groups when picking classes that exist at the highest level.

In summary, the literature suggests BIC and SABIC performed best for class enumeration for both single-level and multilevel mixture models. However, no study explicitly examined which indices perform best when the higher level of nesting is ignored in the context of MGMM. In addition, none of the current widely used fit indices for class enumeration was designed/developed for testing multilevel models. It would be interesting to see if using the individual level N would be sufficient for the more common MGMM. To address this shortcoming in the literature, this study tests the performance of six commonly used model selection indices in MGMM with continuous outcomes, including AIC, CAIC, BIC, SABIC, VLMR, and ALMR, all of which except for CAIC are available in Mplus (L. K. Muthén & Muthén, 1998-2010). The performances of these indices are compared under the correct model specification, which includes the higher level of the nesting to accommodate the clustered data structure, and under the misspecified model, where the higher level is ignored. This study's results can provide insights into the robustness of the model selection indices when a higher level in MGMM is ignored under a variety of design conditions as well as the best practice to use when such misspecification is inevitable.

Table 1. Summary of Recent Studies Examining Performance of Model Selection Fit Indices in Mixture Models.

Reference	Model(s) studied			Indices studied	Better reforming indices
	Name	Design	Type		
Yang (2006)	Latent class analysis	Cross-sectional	Dichotomous	AIC, CAIC, BIC, SABIC, DBIC, SACAIC, HQ-IC, HT-AIC	SABIC
Henson, Reise, and Kim (2007)	Latent variable mixture model	Cross-sectional	Continuous	AIC, CAIC, BIC, SABIC, VLMR, ALMR, CLC, ICL-BIC, NEC, entropy, MST, MKT	SABIC, VLMR/ALMR
Nylund, Asparouhov, and Muthén (2007)	Latent class analysis	Cross-sectional	Dichotomous or continuous	AIC, CAIC, BIC, SABIC, VLMR, NCS, BLRT	BLRT, BIC, SABIC
Tofghi and Enders (2008)	Factor mixture model	Cross-sectional	Categorical		
Morgan, Hodge, and Baggett (2016)	Growth mixture model	Longitudinal	Continuous		
Clark and Muthén (2007)	Growth mixture model	Longitudinal	Continuous	AIC, CAIC, BIC, SABIC, VLMR, ALMR, MST, MKT	SABIC, VLMR/ALMR
Allua (2007)	Latent profile analysis	Cross-sectional	Continuous	AIC, AICc, CAIC, BIC, SABIC, DIC, ICL-BIC, entropy, ALMR	BIC, SABIC, CAIC, ALMR
Lukociene and Vermunt (2010)	Latent class analysis	Cross-sectional, clustered	Dichotomous	AIC, BIC, SABIC	None for simple structure data, BIC, SABIC for complex data structure
Peugh and Fan (2012)	Multilevel factor mixture model	Cross-sectional, clustered	Continuous	AIC, BIC, SABIC, ALMR	BIC, SABIC, ALMR better when only one class in population; AIC, ALMR better when two classes in population.
	Multilevel mixture model	Cross-sectional	Binary	BIC, AIC, AIC3, CAIC, ICOMP, validation log-likelihood	BIC, CAIC, ICOMP under large separation, AIC under low separation and small group size
	Growth mixture model	Longitudinal	Continuous	AIC, CAIC, BIC, DBIC, HQ, HT-AIC, NC-AIC, NBIC, ND-BIC, NHQ, NHT-AIC, Entropy, NEC, CLC, ICL-BIC, NICL-BIC, LMR, ALMR, BLRT	For $k = 1$, CLC, ICL-BIC, and NICL-BIC. For $k = 3$, entropy and HT-AIC for small sample, simpler model; HT-AIC and entropy for more complex model with small sample; BLRT for more complex model with large sample.
Bacci, Pandolfi, and Pennoni (2014)	Latent Markov model	Longitudinal	Categorical	BIC, AIC, AIC3, AICc, CAIC, NEC, NEC1, NEC2, CLC, and ICL-BIC	NEC1 and NEC2 account for all numbers of time points and latent classes tested.

Note. Alphabetized fit indices. AIC = Akaike information criterion; AICc = corrected Akaike's information criterion; AIC3 = A variant of AIC that uses $3p$ as the penalty term; ALMR = adjusted Lo-Mendell-Rubin likelihood ratio test; BIC = Bayesian information criterion; BLRT = bootstrap likelihood ratio test; CAIC = consistent AIC; CLC = classification likelihood information criterion; DBIC = Draper BIC; DIC = deviance information criterion; HQ-IC = Hannan-Quinn information criterion; HT-AIC = Hurvich-Tsai AIC; ICL-BIC = integrated classification likelihood (BIC approximation); ICOMP = Bozdogan's information complexity criterion; LMR = Lo-Mendell-Rubin likelihood ratio test; MKT = multivariate kurtosis test; MST = multivariate skewness test; NBIC = sample size-adjusted BIC; NC-AIC = sample size-adjusted CAIC (aka SACAIC); NCS = traditional chi-square difference test; ND-BIC = sample size-adjusted DBIC; NEC = normalized entropy criterion; NHQ = sample size-adjusted Hannan-Quinn information criterion; NHT-AIC = sample size-adjusted HT-AIC; NICL-BIC = sample size-adjusted ICL-BIC; SABIC = sample size-adjusted BIC (aka NBIC); SACAIC = sample size-adjusted CAIC (aka NC-AIC); VLMR = Vuong-Lo-Mendell-Rubin likelihood ratio test.

Study I: Two-Class Case

Method

Data generation. Data with two known subpopulations under a three-level model (e.g., repeated measures nested within students and students nested within schools) were first generated. The three-level model for data generation is shown below:

Level 1:

$$Y_{ij} = \pi_{0ij} + \pi_{1ij}(\text{Time})_{ij} + e_{ij}, \quad (2a)$$

with

$$e_{ij} \sim N(0, \sigma^2). \quad (2b)$$

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j} \text{subpopulation}_{ij} + r_{0ij}, \quad (2c)$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j} \text{subpopulation}_{ij} + r_{1ij}, \quad (2d)$$

with

$$\begin{bmatrix} r_{0ij} \\ r_{1ij} \end{bmatrix} \sim MVN \left(\mathbf{0}, \mathbf{T}_\pi = \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{bmatrix} \right). \quad (2e)$$

Level 3:

$$\beta_{00j} = \gamma_{00} + \mu_{0j}, \quad (2f)$$

$$\beta_{01j} = \gamma_{01}, \quad (2g)$$

$$\beta_{10j} = \gamma_{10}, \quad (2h)$$

$$\beta_{11j} = \gamma_{11}, \quad (2i)$$

with

$$\mu_{0j} \sim N(0, \tau_{\beta 00}), \quad (2j)$$

where the time variable $(\text{Time})_{ij}$ was centered and had values of $[-1.5, -0.5, 0.5, 1.5]$, and $\text{subpopulation}_{ij}$ was a dichotomized variable with 0 and 1 representing two different subpopulations. The subscript t represents the measurement occasions ($t = 1, 2, 3, 4$), the subscript i represents the individuals ($i = 1 \dots n_j$), and the subscript j represents the clusters ($j = 1 \dots J$). We used four repeated measures for all simulation conditions because (a) previous studies found no significant effect of the number of repeated measures on model estimation, and (b) four waves of repeated measures were most commonly used in both simulation studies (Enders & Tofighi, 2008; Nylund et al., 2007; Tofighi & Enders, 2008) and empirical studies (Khoo, West, Wu, & Kwok, 2006).

In this three-level model, four fixed effect coefficients (i.e., γ_{00} , γ_{01} , γ_{10} , and γ_{11}) and five variances and covariances

of the random effects (i.e., σ^2 , $\tau_{\pi 00}$, $\tau_{\pi 01}$, $\tau_{\pi 10}$, $\tau_{\pi 11}$, $\tau_{\beta 00}$) needed to be specified. The average growth models for the two subpopulations were specified as follows so that Subpopulation A represents a low-start and slow-growing group and Subpopulation B represents a high-start and fast-growing group:

Subpopulation A:

$$\hat{Y}_{ij} = 1.00 + .10 \times (\text{Time})_{ij}, \quad (3a)$$

Subpopulation B:

$$\hat{Y}_{ij} = 2.50 + .60 \times (\text{Time})_{ij}. \quad (3b)$$

Based on the settings presented in Equations 2a and 2b, γ_{00} , γ_{01} , γ_{10} , and γ_{11} were set to 1, 1.5, 0.1, and 0.5, respectively. The residual variance was set to $\sigma^2 = 1.0$.

Design factors. Previous simulation studies have identified some important design factors that may affect the performance of the model selection indices. First, the degree of class separation dramatically impacts enumeration of the correct number of classes; if the generated classes are well-separated, the correct number of classes is more easily identified (Henson et al., 2007; Tofighi & Enders, 2008). Second, the latent class mixing proportions have a substantial impact on class enumeration (Enders & Tofighi, 2008; Henson et al., 2007; Tofighi & Enders, 2008). In unbalanced situations where one latent class has an extremely low mixing proportion (e.g., 7% in Tofighi & Enders, 2008, 10% in Henson et al., 2007), the model is less likely to converge and the class enumeration is less accurate. Third, sample size influences model selection index performance in class enumeration, performing better with larger sample sizes (Henson et al., 2007; Tofighi & Enders, 2008). Given these findings, we manipulated five design factors: degree of separation, conditional intraclass correlation (ICC), number of clusters, cluster size, and latent class mixing proportions.

Degree of separation. We manipulated the magnitude of the \mathbf{T}_π matrix, which includes the within-class variance parameters (see Equation 2e), to produce different degrees of separation. Although some research has implemented algorithms that generate mixture distribution data using data characteristics such as pairwise overlap between classes (e.g., Maitra & Melnykov, 2010; Melnykov, Chen, & Maitra, 2012), we manipulated the magnitude of the \mathbf{T}_π matrix because other research has shown that it is a sufficient method for generating data from a mixture distribution (e.g., Chen et al., 2010). Holding the mean growth factors of the two subpopulations constant, the larger the variation of individual growth trajectories within each subpopulation, the more overlapping and less separated the two subpopulations are. Following Raudenbush and Liu's (2001) criteria,

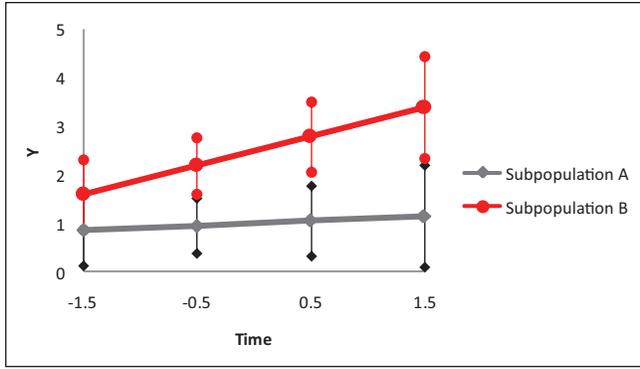


Figure 1. Mean growth trajectories with error bars for small $\mathbf{T}_{\pi 1}$ matrix (medium separation).

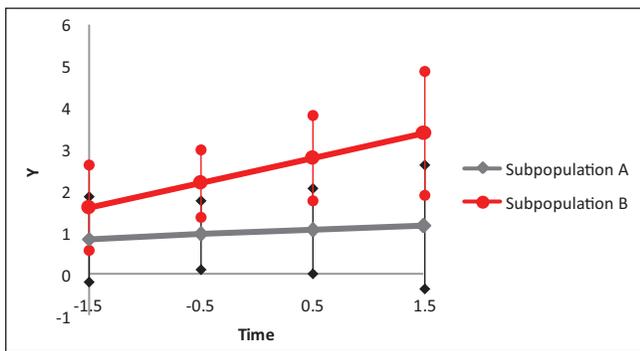


Figure 2. Mean growth trajectories with error bars for medium $\mathbf{T}_{\pi 1}$ matrix (low separation).

we used $\mathbf{T}_{\pi 1} = \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{bmatrix} = \begin{bmatrix} .100 & .025 \\ .025 & .050 \end{bmatrix}$ as small \mathbf{T}_{π} and $\mathbf{T}_{\pi 2} = \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{bmatrix} = \begin{bmatrix} .200 & .050 \\ .050 & .100 \end{bmatrix}$ as medium \mathbf{T}_{π} . As illus-

trated in Figures 1 and 2, with the small $\mathbf{T}_{\pi 1}$ matrix, the two subpopulations only overlap slightly at the first time point but do not overlap at the subsequent time points¹, indicating a medium level of separation. However, with the medium $\mathbf{T}_{\pi 2}$ matrix, the two subpopulations overlap at all four time points, indicating a low level of separation.

Conditional ICC. We selected two levels of conditional ICC, .10 and .20, to represent small clustering effect and medium clustering effect (Hox, 2010). Based on the conditional ICC and \mathbf{T}_{π} matrix, the value of $\tau_{\beta 00}$ was determined by conditional ICC = $\tau_{\beta 00} / (\sigma^2 + \tau_{\pi 00} + \tau_{\beta 00})$. Hence, $\tau_{\beta 00}$ was 0.122 when the conditional ICC was .10 and the \mathbf{T}_{π} matrix was small, 0.133 when the conditional ICC was .10 and the \mathbf{T}_{π} matrix was medium, 0.275 when the conditional ICC was .20 and the \mathbf{T}_{π} matrix was small, and 0.300 when the conditional ICC was .20 and the \mathbf{T}_{π} matrix was medium.

Number of clusters. We considered three levels for the number of clusters: 30, 50, and 80. Recently, Graves and

Frohwerk (2009) systematically reviewed 27 studies using multilevel modeling from five journals devoted to school psychology research and practice. For the 27 studies, the cluster number (e.g., number of schools) had a mean of 28 and a minimum of 17. However, we use 30 as the minimal number of clusters because the multilevel modeling research design literature suggests at least 30 clusters are needed to provide unbiased estimates of fixed and random effects that can be expected to replicate in repeated samples from the same population (Hox, 2010; Kreft & De Leeuw, 1998). We included 50 and 80 as medium and high cluster number levels, which enables our simulation study to mimic applied studies with higher cluster numbers and/or in areas other than school psychology, and allows us to examine the impact of a broader range of cluster numbers and overall sample size combinations on the estimation of the MGMMs.

Cluster size. We selected two levels for cluster size: 20 and 40 individuals per cluster. Based on Graves and Frohwerk's (2009) review, the average cluster size was 44 ($SD = 43$). The level of 40 individuals per cluster was close to the mean cluster size while the level of 20 individuals per cluster was close to the 50th percentile ($n = 24$).

To further justify the sample size conditions, we conducted a literature search in PSYCINFO (from year 2000 to 2011) for empirical studies applying GMM in different substantive areas. We found a total of 171 studies; however, only one recent study used MGMM (i.e., Tobler & Komro, 2010). After removing eight studies with extreme sample sizes, the overall sample size of the remaining 163 studies ranged from 115 to 5,914, with a mean of 969 ($SD = 1,204$). Based on the design factor levels of cluster number (30, 50, and 80) and cluster size (20, 40), the combined/overall sample size in our simulation study covered a wide range (i.e., from 600 to 3,200) that is common in applied studies in social sciences.

Mixing proportion. The mixing proportions of the two subpopulations were set to be balanced or unbalanced. In the balanced situation, mixing proportion was set to 50% and 50% for the two subpopulations. In the unbalanced situation, the mixing proportion was set to 25% for the low-start and slow-growing group and 75% for the high-start and fast-growing group, to mimic a situation in school setting where the majority of students develop their reading skills quickly (Nylund et al., 2007). We did not consider a more extreme unbalanced situation because previous research found that models with extreme population mixture proportions (i.e., 10% or less) of a subpopulation were less likely to converge and the class enumeration was less accurate (Henson et al., 2007; Tofighi & Enders, 2008).

In summary, the simulation used a 2 (degree of separation: low or medium) \times 2 (cluster size: 20 or 40 cases) \times 3 (number of clusters: 30, 50, or 80 clusters) \times 2 (mixing proportions: 50%:50% or 75%:25%) \times 2 (ICC: .10 or .20) factorial design to generate the data. A total of 500 replications were generated for each condition using the SAS 9.2 Proc

Table 2. Example of Relationship Between AIC Difference and Akaike Weight.

Δ_3	AW_1	AW_2	AW_3
4	0.11	0.79	0.11
3	0.10	0.74	0.16
2	0.09	0.67	0.24
1	0.08	0.57	0.35
0.1	0.06	0.48	0.46

Note. AIC = Akaike information criterion; AW = Akaike weight; subscripts 1, 2, and 3 refer to one-class, two-class, three-class models, respectively; $\Delta_3 = AIC_3 - AIC_2$.

IML procedure, yielding a total of 24,000 datasets (500 datasets \times 48 conditions). For each replication, six different models—that is, 2 model specifications (true/misspecified) \times 3 different class solutions with different numbers of classes (1, 2, and 3) = 6 models—were fitted using the Mplus 6.1 Mixture routine (L. K. Muthén & Muthén, 1998-2010); analyses were conducted using the MLR (MLR is a robust maximum likelihood estimator that uses a Huber-White sandwich approach to adjust standard errors for non-normality) estimator. The number of initial stage random start values was initially set to 50 and the number of final stage optimizations was set to 10. If there were any convergence issues during analysis, the number of random starts was increased.

Analysis. All 24,000 datasets had converged results for fitted models and a reasonable number of observations in each latent class (i.e., class size no less than 6% of the total sample size). The performance of a model selection index was measured by the proportion of replications in which the index retrieved the correct number of classes.

Criterion for determining success in class enumeration. For AIC, differences less than 2 suggest no credible evidence as to which model is better, differences between 2 and 4 weak evidence, differences between 4 and 7 definite evidence, differences between 7 and 10 strong evidence, and differences larger than 10 very strong evidence (Burnham & Anderson, 2002). We adopted the cutoff value of 4 points; specifically, for AIC to select the correct two-class solution, it had to satisfy two requirements. First, the two-class AIC had to be smaller than the one-class and three-class AIC. Second, the two-class AIC had to be 4 or more points smaller than the one-class AIC. Based on the K-L information theory, AIC differences can be converted to Akaike weight w_i^2 for each model tested in a set (Burnham & Anderson, 2002, 2004). The Akaike weight of a model can be interpreted as the probability associated with the model. Table 2 shows the Akaike weights for the one-class, two-class, and three-class solutions when AIC_{2class} is 4 points less than AIC_{1class} . We can see that using our rule, the two-class solution always has the highest probability, even when the difference between the two-class and three-class AIC is small.

For BIC, differences less than 2 suggest weak evidence, differences between 2 and 6 positive evidence, differences between 6 and 10 strong evidence, and differences larger than 10 very strong evidence (Raftery, 1996). The difference in BIC can also be converted into relative probability. For instance, if the two-class BIC is 2 points less than the one-class BIC, the Bayes factor for a two-class model against a one-class model is 3 (i.e., $B_{21} \approx \exp((BIC_1 - BIC_2) / 2)$), and the posterior probability associated with the two-class solution is .75 (Raftery, 1996). Hence, we adopted the cutoff value of 2 points for the BIC.

For CAIC and SABIC, no guideline has been proposed before. We used 2 as their cutoff values because they are similar to BIC in computation.

As previously mentioned, different rules were used to select best-fitting models for VLMR and ALMR. Because these two indices use a significance test, when testing the k -class model ($k > 1$), there are two possible decisions. When the p value of these indices is equal or smaller than .05, the k -class model would be selected over the $(k - 1)$ -class model; when the p value is larger than .05, there is a lack of evidence for significant improvement and therefore the more parsimonious model—that is, $(k - 1)$ -class model—is selected (Lo et al., 2001). The procedure for determining the best model (see Figure 3) begins with the two-class model, and continues until a comparison of the k -class model, and the $(k + 1)$ -class identify the k -class solution as the better model. We stopped at the three-class model because our interest was in whether the correct number of classes was identified and any decision other than the two-class solution would be considered a wrong decision.

Difference between true and misspecified models. We examined how differently the indices performed under the true model (i.e., consider the higher level) and the misspecified model (i.e., ignoring the higher level). The accuracy of each index under the true and misspecified models was compared and differences were computed.

Impact of the design factors. ANOVAs were conducted to determine the impact of the five design factors on the fit indices' class enumeration accuracy. The percentage of correct model identification was the analysis outcome (e.g., the outcome value is 90% if two-class model was selected for 450 out of 500 replications). Eta-squared (i.e., $\eta^2 = SS_{Effect} / SS_{Total}$) was computed as the effect size indicator. With a balanced design, the ANOVA results are expected to be robust regardless of the sampling distributions of the various statistics, some of which can be expected to be non-normal (Glass & Hopkins, 1996).

Results

Overall fit index performance. Figure 4 shows the average percentages of one-class, two-class, and three-class (or more) models identified by AIC, CAIC, BIC, SABIC, VLMR, and ALMR for all 24,000 datasets under both true and misspecified

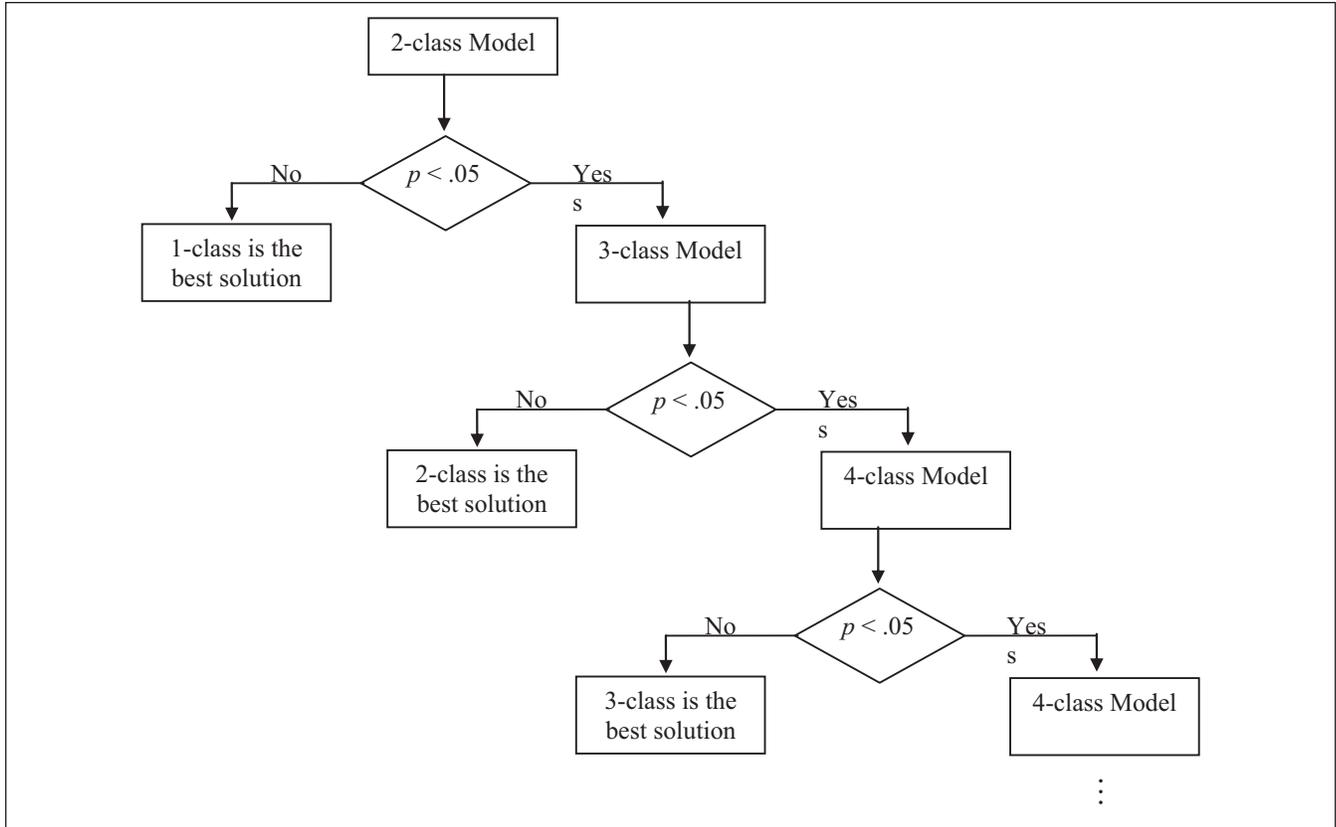


Figure 3. VLMR/ALMR decision flowchart.

Note. VLMR = Vuong–Lo–Mendell–Rubin likelihood ratio test; ALMR = adjusted Lo–Mendell–Rubin likelihood ratio test.

models. As shown in the figure, all model selection indices correctly identified the two-class solution (i.e., the correct solution) in most replications regardless of model specification (i.e., correctly specified or misspecified model). For the true model, BIC had the highest percentage of correct classification (98%), followed by SABIC and CAIC (97%), ALMR (90%), VLMR (89%), and AIC (76%). For the misspecified model, SABIC had the highest percentage of correct classification (91%), followed by BIC (82%), ALMR (81%), VLMR (80%), CAIC (78%), and AIC (66%). The difference in classification accuracy between true and misspecified models was 6%, 9%, 9%, 9%, 16%, and 20% for SABIC, VLMR, ALMR, AIC, BIC, and CAIC, respectively.

Under the true model, the proportion of inconclusive classification ranged from approximately 20% for AIC, to almost 0 for CAIC, BIC, and SABIC. More inconclusive classification appeared under the misspecified model, with AIC again having the highest proportion, followed by BIC and CAIC, and then SABIC. CAIC and BIC had a tendency to underextract the number of classes under the misspecified models, whereas AIC, VLMR, and ALMR tended to overextract the number of classes under both the true and misspecified models.

Effect of design factors. ANOVA results indicated that ICC, cluster number, and cluster size were the three most important factors. Table 3 shows the performance of the six model

selection indices under both the true and false models as well as the difference between the true and misspecified models collapsed over the three factors.

ICC had a negative impact on all six model selection indices under the false model. The average accuracy of class enumeration decreased as ICC increased from .1 to .2 (η^2 ranged from .12 to .16).

In general, cluster number and cluster size had positive effects on the accuracy of the IC indices. As cluster number and cluster size increased, the accuracy of class enumeration increased ($\eta^2_{\text{Clusternumber}}$ ranged from .07 to .43; $\eta^2_{\text{Clustersize}}$ ranged from .07 to .30), except that cluster size had no substantial impact on AIC under the misspecified model. However, for VLMR and ALMR under the true model, cluster number affected classification accuracy negatively ($\eta^2_{\text{Clusternumber}} = .14$ and .09 for VLMR and ALMR, respectively), and cluster size had no substantial effect. Specifically, the accuracy of VLMR and ALMR did not change substantially when cluster number increased from 30 to 50, but when cluster size changed from 50 to 80, the classification accuracy decreased about 4% on average.

In addition, increasing T_{π} matrix decreased the accuracy for CAIC, BIC, VLMR, and ALMR under both the true and misspecified models, and SABIC under the misspecified model only (η^2 ranged from .06 to .14). As the mixing proportions changed from balanced to unbalanced for the two

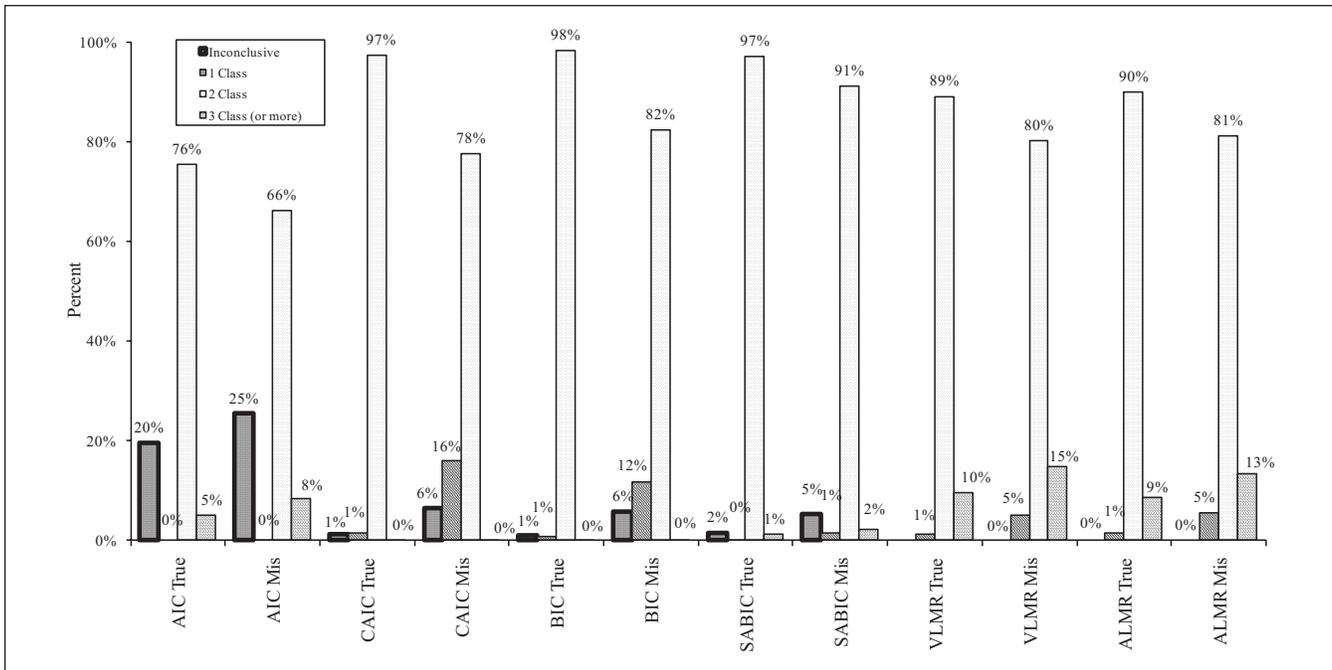


Figure 4. Percentage of one-, two-, three-class (or more) models and inconclusive classification identified by AIC, CAIC, BIC, SABIC, VLMR, and ALMR under true and misspecified models. Note. Three-class (or more): three-class for AIC/CAIC/BIC/SABIC, three-class or more for VLMR/ALMR. AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC = sample size-adjusted BIC; VLMR = Vuong-Lo-Mendell-Rubin likelihood ratio test; ALMR = adjusted Lo-Mendell-Rubin likelihood ratio test.

subpopulations, the accuracy of class enumeration increased for VLMR and ALMR under the true model, and for CAIC, BIC, and SABIC under the misspecified model, but decreased for AIC under the true model (η^2 ranged from .06 to .47).

Study 2: Three-Class Case

Method

In Study 2, we examined the performance of the previously mentioned model selection indices in a three-subpopulation MGMM. According to Tofghi and Enders's (2008) review, three-class models are also common in published studies using GMM. In addition, increased number of subpopulations might increase the difficulty of separating latent classes. Therefore, Study 2 can help determine whether fit index performance remains consistent in more complicated situations. The design conditions and analysis procedure remained mostly similar as those in Study 1 with a few modifications, which are described in the following sections.

Data generation. In Study 2, data with three known subpopulations under a three-level model were first generated. The Level 1 model was the same as in Study 1. The Level 2 and Level 3 models were as follows:

Level 2:

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}D1_{ij} + \beta_{02j}D2_{ij} + r_{0ij}, \quad (4a)$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}D1_{ij} + \beta_{12j}D2_{ij} + r_{1ij}, \quad (4b)$$

with

$$\begin{bmatrix} r_{0ij} \\ r_{1ij} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma}_{\pi} = \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} \\ \tau_{\pi 10} & \tau_{\pi 11} \end{bmatrix} \right). \quad (4c)$$

Level 3:

$$\beta_{00j} = \gamma_{00} + \mu_{0j}, \quad (4d)$$

$$\beta_{01j} = \gamma_{01},$$

$$\beta_{02j} = \gamma_{02}, \quad (4e)$$

$$\beta_{10j} = \gamma_{10}, \quad (4f)$$

$$\beta_{11j} = \gamma_{11},$$

$$\beta_{12j} = \gamma_{12}, \quad (4g)$$

with

$$\mu_{0j} \sim N(0, \tau_{\beta 00}), \quad (4h)$$

Table 3. Study 1, Study 2, and Study 3 Results Collapsed Over ICC, Cluster Number, and Cluster Size.

		Performance of fit indices										Difference between models																	
Design conditions		Performance of fit indices										Difference between models																	
Study	ICC	Cluster number	Cluster size	n	AIC					SABIC					ALMRf		ALMRt		CAIC		BIC		SABIC		VLMR		ALMR		
					AICt	AICf	CAICt	CAICf	BICt	BICf	SABICt	SABICf	VLMRt	VLMRf	ALMRf	ALMRt	ALMRf	ALMRt	AIC	CAIC	BIC	SABIC	VLMR	ALMR					
Study 1	0.1	30	20	600	73	66	87	55	91	66	93	85	88	80	88	80	88	80	7	32	25	8	8	8	8	8	8	8	8
		40	40	1,200	77	72	100	91	100	94	98	97	93	84	94	84	94	85	5	9	6	1	9	9	9	9	9	9	9
		50	20	1,000	75	70	98	83	99	88	97	95	91	83	92	85	85	85	5	15	11	2	8	7	8	7	8	7	7
	0.2	80	40	2,000	77	72	100	99	100	99	98	98	89	84	90	85	85	85	5	1	1	0	5	5	5	5	5	5	5
		20	20	1,600	76	72	100	97	100	98	99	98	87	85	88	87	85	88	87	3	3	2	1	2	1	2	1	2	1
		40	40	3,200	77	71	100	100	100	100	99	99	86	82	87	83	82	87	83	6	0	0	0	4	4	4	4	4	4
Study 2	0.1	30	20	600	71	55	87	24	92	34	92	67	88	65	88	64	88	64	16	63	58	26	23	24	24	24	24	24	24
		40	40	1,200	76	62	100	66	100	73	98	87	93	80	94	81	94	81	14	33	27	11	13	13	13	13	13	13	13
		50	20	1,000	73	63	99	52	99	61	96	84	91	77	92	77	92	77	11	47	38	12	15	15	15	15	15	15	15
	0.2	80	40	2,000	77	62	100	88	100	92	98	94	92	81	93	83	93	83	15	12	8	4	11	10	10	10	10	10	10
		20	20	1,600	77	68	100	79	100	85	99	95	87	82	89	84	89	84	9	21	15	4	5	5	5	5	5	5	5
		40	40	3,200	78	63	100	98	100	99	99	97	85	81	87	83	87	83	15	2	1	2	4	4	4	4	4	4	4
Study 3	0.1	30	21	630	52	57	92	79	96	88	85	87	82	79	82	79	82	79	-5	13	7	-2	3	2	2	2	2	2	
		42	42	1,260	39	42	100	100	99	100	84	85	94	74	94	76	94	76	-3	0	-1	-1	20	19	19	19	19	19	
		80	21	1,680	33	38	99	100	99	99	84	85	83	71	85	73	85	73	-5	0	0	-1	11	12	12	12	12	12	
	0.2	42	42	3,360	16	18	99	98	97	97	72	78	67	54	69	56	69	56	-2	1	0	-6	13	14	14	14	14	14	
		21	21	630	53	51	93	49	96	64	85	82	83	72	81	72	81	72	2	44	31	3	11	9	9	9	9	9	
		42	42	1,260	40	42	100	95	100	97	86	86	94	74	94	77	94	77	-2	5	2	0	19	18	18	18	18		
0.2	80	21	1,680	34	39	99	100	99	99	85	86	85	74	87	76	87	76	-6	0	0	-1	11	11	11	11	11	11		
	42	42	3,360	13	19	98	98	97	97	72	81	66	57	69	60	69	60	-6	0	0	-9	9	9	9	9	9	9		
	20	20	600	39	25	100	97	99	97	77	62	97	85	97	85	97	85	14	2	2	15	12	12	12	12	12	12		
40	40	3,200	39	29	100	100	100	100	99	95	96	83	83	96	84	96	84	10	0	0	3	13	13	13	13	13	13		

Note. ICC = intraclass correlation; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC = sample size-adjusted BIC; VLMR = Vuong-Lo-Mendell-Rubin likelihood ratio test; ALMR = adjusted Lo-Mendell-Rubin likelihood ratio test. The "t" represents true model and "f" represents misspecified model. Numbers for performance and difference are in percentage numbers. Difference between models = true model accuracy - misspecified model accuracy.

where $D1_{ij}$ and $D2_{ij}$ were dummy variables to represent the three different subpopulations (i.e., $D1_{ij} = 1$ and $D2_{ij} = 0$ for Subpopulation A; $D1_{ij} = 0$ and $D2_{ij} = 1$ for Subpopulation B; and $D1_{ij} = 0$ and $D2_{ij} = 0$ for Subpopulation C).

The average growth models for the three subpopulations were specified so that the two subpopulations in Study 1 remained the same as in Equations 2a and 2b, and a third subpopulation with a high start but a decelerating mean growth trajectory (i.e., Equation 3k) was added: Subpopulation C:

$$\hat{Y}_{ij} = 3.00 - .30 \times (\text{Time})_{ij}. \quad (4i)$$

We chose the growth pattern for Subpopulation C based on a review by Tofighi and Enders (2008). The shape of estimated growth trajectories usually includes three classes: (a) a “zero class” of individuals with low and stable levels of some problem behaviors (i.e., Subpopulation A), (b) an “accelerating class” with low start but increasing number of problems (i.e., Subpopulation B), and (c) a “decelerating class” with higher start but decreasing number of problems (i.e., Subpopulation C).

Based on the settings presented in Equations 2a, 2b, and 3k and the coding of the dummy variables, γ_{00} , γ_{01} , γ_{02} , γ_{10} , γ_{11} , and γ_{12} were set to 3, -2, -0.5, -0.3, 0.4, and 0.9, respectively. Hence, the intercepts for each subpopulation were 1.00 (I1), 2.50 (I2), and 3.00 (I3), and the slopes were 0.10 (S1), 0.50 (S2), and -0.30 (S3), respectively.

The three subpopulations’ mixing proportions were fixed to be balanced, with 33.33% for each subpopulation, because the effect of mixing proportions was not substantial according to Study 1’s findings. The cluster size was set to be 21 and 42 for the ease of assigning equal number of individuals into three subpopulations during data generation through SAS (version 9.2; SAS Institute Inc., 2002-2008). Because the impact of cluster number has been clearly shown in Study 1, we only adopted the low and high levels for this design factor (i.e., 30 and 80) and omitted the intermediate level (i.e., 50) to reduce the total number conditions. The variances and covariance of the random effects as well as the number of repeated measures were the same as in Study 1. The mean growth trajectories of the three subpopulations and the level of separation under the small and medium \mathbf{T}_π matrix are illustrated in Figures 5 and 6.

In summary, the simulation used a 2 (magnitude of the \mathbf{T}_π matrix: small or medium) \times 2 (number of participants per cluster: 21 or 42 cases) \times 2 (number of clusters: 30 or 80 clusters) \times 2 (ICC: .10 or .20) factorial design to generate the data. Similar to Study 1, 500 replications were generated for each condition yielding a total of (500 datasets \times 16 conditions) 8,000 datasets.

Analyses and Results

The analysis procedures and observed outcomes were the same as those for Study 1. The performance of the six fit

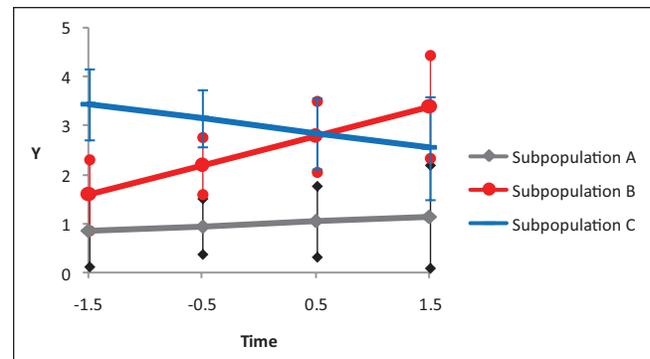


Figure 5. Mean growth trajectories with error bars for small $\mathbf{T}_{\pi 1}$ matrix (medium separation).

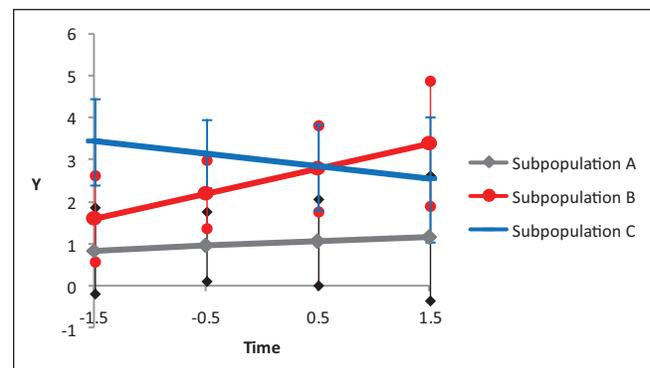


Figure 6. Mean growth trajectories with error bars for medium $\mathbf{T}_{\pi 2}$ matrix (low separation).

indices had several similarities with those in Study 1. Therefore, we only highlighted patterns that are different.

Overall fit index performance. Results including the percentages of correct and incorrect classifications by the six model selection indices for both the true and misspecified models are presented in Figure 7. Because Study 2 contained fewer design conditions than Study 1, only results from design conditions similar across two studies were used for comparison. AIC had a much lower class enumeration accuracy (below 40%) compared with that in Study 1 (above 65%) for both true and misspecified models. CAIC and BIC still performed well in true models; furthermore, their enumeration accuracy under the misspecified models improved by 13% and 11%, respectively, compared with that in Study 1. Compared with results from Study 1, SABIC’s accuracy decreased by 16% under the true model and 7% under the misspecified model, with the overextraction rate increasing by 6% and 4%, respectively. The accuracy of VLMR and ALMR dropped by 7% under the true model and 10% under the misspecified model compared with that in Study 1.

In summary, under the misspecified model, BIC and CAIC had higher percentages of correct class enumeration (i.e., 93% and 90%) than SABIC (84%), followed by VLMR

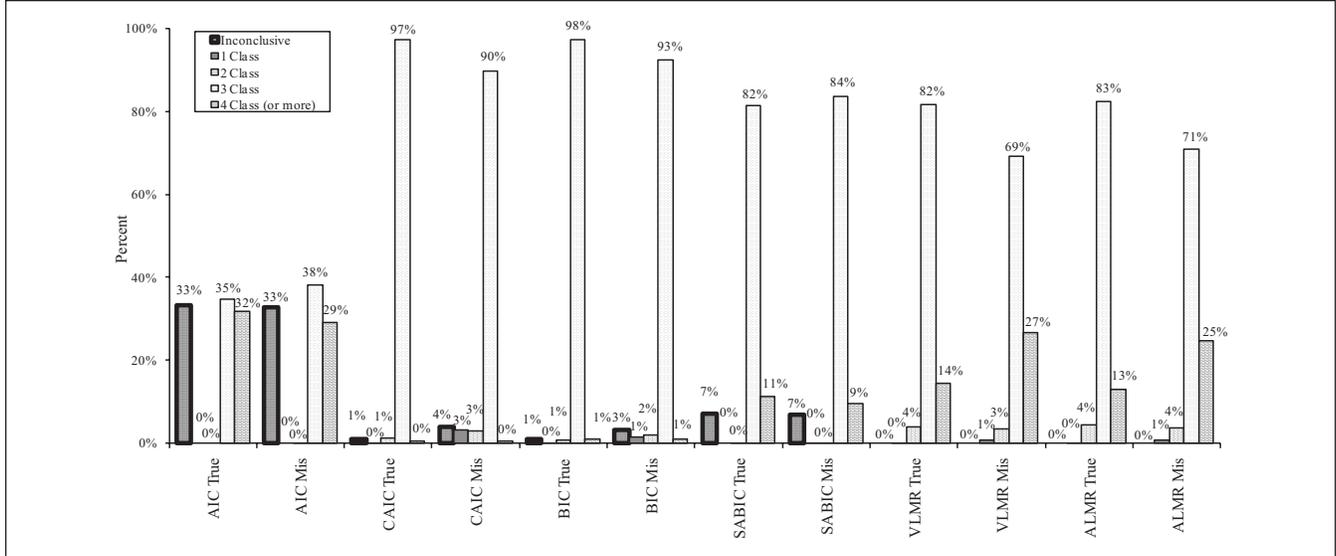


Figure 7. Percentage of one-, two-, three-, four-class (or more) models and inconclusive classification identified by AIC, CAIC, BIC, SABIC, VLMR, and ALMR under true and misspecified models.

Note. Four-class (or more): four-class for AIC/CAIC/BIC/SABIC, four-class or more for VLMR/ALMR. AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC = sample size-adjusted BIC; VLMR = Vuong–Lo–Mendell–Rubin likelihood ratio test; ALMR = adjusted Lo–Mendell–Rubin likelihood ratio test.

and ALMR (69% and 71%). The difference in classification accuracy between true and misspecified models was -2% , -3% , 5% , 8% , 12% , and 12% for SABIC, AIC, BIC, CAIC, ALMR, and VLMR, respectively.

Effect of design factors. The effects of the design factors remained mostly similar as what was found in Study 1; therefore, only a few different patterns are highlighted here. For AIC, cluster number and cluster size affected the classification accuracy negatively ($\eta^2_{\text{clusternumber}} = .46$ and $.38$, $\eta^2_{\text{clustersize}} = .24$ and $.26$ for true and misspecified models, respectively). As cluster size and number increased, the class enumeration accuracy for all AIC decreased. The accuracy of SABIC also decreased significantly (i.e., overall 13% for true model and 5% for misspecified model) when sample size was at the highest level. In addition, ICC did not affect the accuracy of SABIC, VLMR, and ALMR under the false model.

Study 3: Complex Model in Two-Class Case

In the first two simulation studies, the within-class covariance structure (i.e., the \mathbf{T}_π matrix) was identical across latent classes. In reality, however, this might not be true. In this simulation, we examined the effect of having different \mathbf{T}_π matrices across classes on model selection index performance. We generated two-class multilevel data with a small \mathbf{T}_π matrix

$$(\mathbf{T}_\pi = \begin{bmatrix} \tau_{\pi00} & \tau_{\pi01} \\ \tau_{\pi10} & \tau_{\pi11} \end{bmatrix} = \begin{bmatrix} .100 & 0 \\ 0 & .050 \end{bmatrix}) \text{ for Population A and a}$$

$$\text{medium } \mathbf{T}_\pi \text{ matrix } (\mathbf{T}_\pi = \begin{bmatrix} \tau_{\pi00} & \tau_{\pi01} \\ \tau_{\pi10} & \tau_{\pi11} \end{bmatrix} = \begin{bmatrix} .200 & 0 \\ 0 & .100 \end{bmatrix}) \text{ for}$$

Population B. The analysis procedure was similar as that of Study 1. Note that now class-specific variances were estimated for each model fitted.

As expected, there were more convergence issues as the model became more complex. A total of 33 replications were excluded from further analysis due to nonconvergence or local solutions for the three-class MGMM or GMM models. We highlighted the results that are different from Study 1 in the following section (see Table 3). AIC’s accuracy decreased substantially, whereas CAIC, BIC, VLMR, and ALMR became more accurate as the model became more complex. SABIC performed similarly as in Study 1, except its accuracy decreased under the small sample size condition under the true model.

Discussion and Conclusion

The current study examined model selection index performance in class enumeration for MGMMs when the top level of nesting was ignored. As expected, under the correctly specified MGMM, the indices’ performance was mostly consistent with the findings in previous studies. BIC and CAIC had the highest class enumeration accuracy, followed by SABIC, ALMR, and VLMR, with ALMR and VLMR performing similarly. AIC had the lowest enumeration accuracy. SABIC, ALMR, VLMR, and AIC tended to overextract the number of classes when there were three subpopulations and less clear-cut class separation. When the highest data level

was ignored and a single-level GMM was fitted to the data, the classification accuracy of all model selection indices decreased compared with the accuracy under the true model. We discuss the impact of design factors, Bonferroni correction with VLMR/ALMR, and the implications for researchers in the following section.

Impact of Design Factors

The impact of ICC. The ICC had the biggest impact on the difference between the true and misspecified models. An important advantage of using multilevel models is that by modeling the higher level nesting structure (e.g., schools), variation in the individual growth trajectories can be decomposed into within- and between-organization components (Raudenbush & Bryk, 2002). As shown in some previous studies on multilevel analysis (e.g., Meyers & Beretvas, 2006; Moerbeek, 2004), ignoring the highest level data structure results in the redistribution of the variance from the ignored level (or the organization/school level) to the adjacent level (i.e., the individual/student level). A similar variance redistribution mechanism has been found in MGMM (e.g., the overestimation of $\tau_{\pi 00}$, Chen et al., 2010). The reduction in class enumeration accuracy is therefore likely the result of the redistributed higher level variance, which can increase the variation of the individual growth trajectories, potentially increasing overlap between different latent classes.

The impact of degree of separation. The magnitude of the \mathbf{T}_{π} matrix is related to the variation within each latent class and thereby the degree of separation of different latent classes. It is not surprising that as the magnitude of \mathbf{T}_{π} matrix increased (i.e., latent classes became less separated), the class enumeration accuracy under both the true and the misspecified models decreased. Furthermore, model selection index performance deteriorated more under the misspecified model than under the true model when the magnitude of \mathbf{T}_{π} increased. This indicates that model selection indices are more sensitive to class separation when the nesting structure is ignored and thus even less likely to identify the correct number of classes when class separation is less clear.

The impact of sample size. Sample sizes affected the indices in different ways. Our finding is consistent with previous findings that CAIC and BIC tend to select the correct class solution more frequently as sample size increases under the true model (Nylund et al., 2007). However, the impact of small sample size on CAIC and BIC under the misspecified GMM is much more serious than its impact under the true model. Previous research on two-level GMM found that BIC and CAIC tended to underextract the number of classes under small sample sizes even when the model was correctly specified. It is possible that the N adjustment is too strong when N is small or \ln is not the optimal function. However,

the performances of SABIC (with smaller penalty on N), VLMR, and ALMR were less affected by small sample size conditions. The performance gaps between the true and the misspecified model for SABIC, VLMR, and ALMR were much smaller than those for CAIC and BIC. Therefore, when sample size was small, SABIC, VLMR, and ALMR were generally more accurate than CAIC and BIC under the misspecified model.

As shown in Equations 1a to 1d, different sample size (N) functions are used in the penalty term for different IC indices. Figure 8 illustrates how class enumeration decisions made by different IC change with sample size based on the log-likelihood values from an empirical data analysis. Suppose we are comparing the three-class solution and the four-class solution. The difference between IC_{4class} and IC_{3class} is

$$\begin{aligned} \Delta IC_{4class-3class} &= (-2\log L_4 + \text{Penalty}_4) - (-2\log L_3 + \text{Penalty}_3) \\ &= (\text{Penalty}_4 - \text{Penalty}_3) - (2\log L_4 - 2\log L_3). \end{aligned}$$

Let $\Delta \text{Penalty} = \text{Penalty}_4 - \text{Penalty}_3$ and $\Delta 2\log L = 2\log L_4 - 2\log L_3$. Figure 8 plotted $\Delta \text{Penalty}$ and $\Delta 2\log L$ against sample size ($1 \leq N \leq 10^4$). The $\Delta 2\log L$ was different under the MGMM and the GMM models (see Figure 8 for the two curved lines). We can see that $\Delta 2\log L$ was positive and increased as the sample size increased. The $\Delta \text{Penalty}$ was different depending on the particular IC index (see Figure 8 for the four straight lines); however, it was the same under MGMM and GMM models because the difference in the number of parameters (i.e., p) between the four- and three-class solutions was the same under MGMM and GMM. We can see that $\Delta \text{Penalty}$ increased as sample size increased for CAIC, BIC, and SABIC whereas the $\Delta \text{Penalty}$ for AIC was a constant despite the sample size increase. When a $\Delta \text{Penalty}$ line intersects a $\Delta 2\log L$ line, it means that $\Delta IC_{4class-3class}$ is zero and the four-class solution is as good as the three-class solution (the corresponding sample size is the cutoff sample size). When the $\Delta \text{Penalty}$ line is above the $\Delta 2\log L$ line, it means that $\Delta IC_{4class-3class}$ is greater than zero and the three-class solution is better. Conversely, if the $\Delta \text{Penalty}$ line is below the $\Delta 2\log L$ line, it means that $\Delta IC_{4class-3class}$ is smaller than zero and the four-class solution is better. Apparently, the cutoff sample size is dependent upon the likelihood values of the tested models and the corresponding number of estimated parameters (i.e., model complexity).

Figure 8 shows that when sample size is large enough, all indices will favor the four-class solution. Note that SABIC will also favor the four-class solution when sample size is very small because it has two crossing points. The range of sample size in which the IC indices will select the three-class solution (i.e., when the $\Delta \text{Penalty}$ line is above the $\Delta 2\log L$ line) is the largest for CAIC, followed by BIC, SABIC, and AIC. In other words, the range of sample size in which the IC indices will favor the four-class solution is the largest for

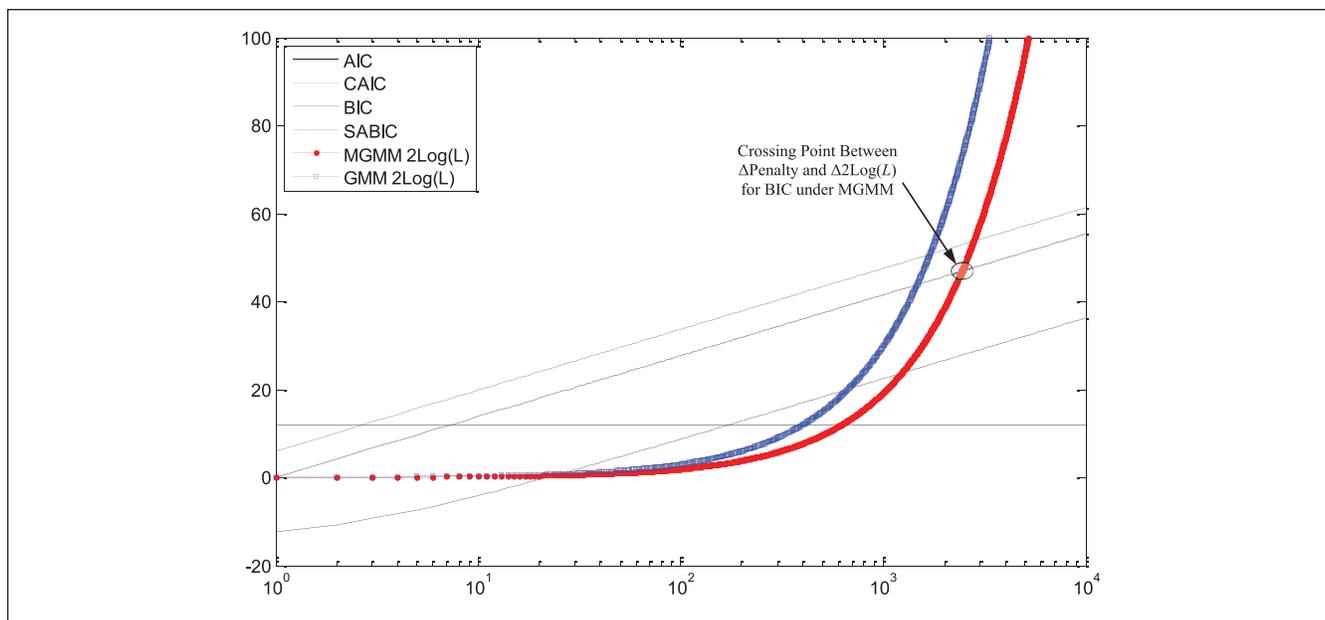


Figure 8. Example of the effect of sample size on class enumeration decisions by IC indices.

Note. IC = information criterion; AIC = Akaike information criterion; CAIC = consistent AIC; BIC = Bayesian information criterion; SABIC = sample size-adjusted BIC; MGMM = multilevel growth mixture model; GMM = growth mixture model.

AIC, followed by SABIC, BIC, and CAIC. This can explain several findings of our simulation studies. First, AIC tended to overextract classes most and SABIC had some overextraction but much less than AIC, whereas BIC and CAIC were more conservative and had less overextraction. This caused the accuracy of AIC and SABIC to decrease with the sample size increase in both studies. Second, SABIC could also overextract when sample size was small because it had two crossing points. This caused the accuracy of SABIC to be lower compared with BIC and CAIC under the smaller sample size condition in Study 2. Third, the range of sample size for an IC index to choose the four-class solution is larger under GMM than under MGMM. This explained why overextraction of classes was more likely to occur under GMM than MGMM.

The MGMM examined in this study was strategically specified to have only one Level 3 random effect, which captures the intercept variance between cluster/organization means. Under specifications where the number of parameters at Level 3 increases (e.g., by including a slope residual variance, latent classes, or fixed effects at the organization level), $\Delta\text{Penalty}$ would remain the same for true and misspecified MGMM; however, these additions may change the amount and shape of the Level 2 variance distribution, which could impact the number of within-cluster classes. Regarding the current study, these Level 3 additions can also impact the accuracy of class enumeration (Chen et al., 2010).

Bonferroni Corrections With VLMR/ALMR

One reason for VLMR/ALMR's overextracting the number of classes might be increased Type I error rate resulting from

multiple testing for the same dataset. In previous research, no correction for α was used and the effect of using corrected α is unknown (B. Muthén, personal communication, March 8, 2011). Therefore, we have examined the accuracy of VLMR and ALMR after applying Bonferroni correction to α in both studies. Based on the number of VLMR/ALMR from each study, we used $\alpha = .05 / 2 = .025$ for Study 1 and $\alpha = .05 / 3 = .167$ for Study 2. We found that the overall accuracy of VLMR and ALMR improved under both the true and misspecified models (improvement ranged from 1% to 10%). Furthermore, the difference between the nonadjusted and adjusted VLMR/ALMR increased as the cluster number and cluster size increased. By using the Bonferroni correction to α , VLMR/ALMR is less likely to overextract the number of classes, an improvement especially noticeable in large sample size conditions, in misspecified models, and when there were more than two subpopulations in data generation. There was no appreciable difference between the nonadjusted and adjusted VLMR/ALMR under the small sample size condition (i.e., $N = 600$ or 630) and for the true model. Therefore, Bonferroni correction seems more appropriate for data with large sample sizes ($N \geq 1,000$) only; besides, when the higher level nesting structure is ignored, the Bonferroni correction would be especially useful to achieve higher accuracy for VLMR/ALMR.

Implications for Researchers

Techniques such as GMM and MGMM that are rooted in structural equation modeling can be viewed as model-building techniques in which researchers start with a simpler model

Table 4. Better Performing Indices Under Small and Large N by Type of Model.

Type of model	Small N	Large N
MGMM		
Two class, same variance	SABIC, BIC, CAIC/VLMR/ALMR	BIC/CAIC, SABIC, VLMR/ALMR
Three class, same variance	BIC, CAIC, VLMR/ALMR, SABIC	BIC, CAIC
Two class, different variance	BIC, CAIC, VLMR/ALMR	BIC, CAIC, SABIC, VLMR/ALMR
GMM		
Two class, same variance	SABIC, VLMR/ALMR	SABIC, BIC, CAIC, VLMR/ALMR
Three class, same variance	SABIC, VLMR/ALMR	BIC, CAIC, SABIC
Two class, different variance	BIC/CAIC, VLMR/ALMR	BIC/CAIC, SABIC, VLMR/ALMR

Note. MGMM = multilevel growth mixture model; SABIC = sample size-adjusted BIC; BIC = Bayesian information criterion; CAIC = consistent AIC; VLMR = Vuong–Lo–Mendell–Rubin likelihood ratio test; ALMR = adjusted Lo–Mendell–Rubin likelihood ratio test; GMM = growth mixture model; AIC = Akaike information criterion.

and build up to a more complex one (Kline, 2010). The results of the current study can aid researchers in this model-building process because they demonstrate the importance of accounting for nested data structure and provide information on which fit indices can most accurately identify the appropriate model when using MGMM. Based on our findings, we strongly recommend that researchers accommodate multi-level structure by using MGMMs, especially when the sample size is relatively small, the ICC is relatively large, or both. Table 4 summarizes the top performing indices under different conditions.

In general, when the sample size is small, SABIC is preferred, whereas when the sample size is large, BIC and CAIC are preferred. Note that the sample size is relative to ICC. For example, 1,200 may be considered as a large sample size when ICC is .1, but small when ICC is .2.

The cutoff value of 2 is a reasonable value to use for BIC, CAIC, and SABIC. In the process of finding the best fit model, when the decrease in the indices' values (especially SABIC) becomes less than 2, researchers can stop fitting more complicated models (i.e., models with more classes) and select the model with the second lowest IC index.

VLMR and ALMR can be used as references when sample size is small (i.e., $N \leq 630$), but should be used with caution when sample size is large due to their tendency to overextract classes. Bonferroni correction can help improve VLMR/ALMR's accuracy for data with large sample size ($N \geq 1,000$).

Under more complex models, we recommend using CAIC, BIC, VLMR, and ALMR for MGMM; SABIC is only suitable for large sample size conditions. For misspecified models, CAIC and BIC are best, SABIC is appropriate for large sample size conditions, and VLMR and ALMR can also be used, but tend to overextract classes.

It is important to note that these fit index recommendations are based on the simulation conditions used in this study. Readers should be cautious when applying our findings to models and conditions that are very different from those studied in this article.

Limitations and Future Directions

Accurate class enumeration is perhaps the greatest challenge in mixture modeling and much is still unknown about highly complex mixture models such as the MGMM. That said, there are some limitations to the current study that raise caution regarding generalizing the results. First, the residual variance of the slope at the organization level was constrained to zero. While this specification will be common for empirical studies because organization level slope variation tends to be small, which may result in convergence problems (Palardy & Vermunt, 2010), including the organization level slope random effect and other Level 3 parameters may impact class enumeration.

Second, while this study focuses on the particular model misspecification of ignoring the highest level of the hierarchical data, there are other types of model misspecifications and assumption violations that may affect class enumeration. For instance, misspecifying the shape of the growth trajectory or violating the multivariate normality assumption for the repeated measures can effect class enumeration (Bauer & Curran, 2003, 2004). Additional research is needed to address these issues.

Third, the current study was conducted under frequentist estimation using the maximum likelihood estimator. Recently, Bayesian estimation for multilevel models has been gaining attention due to its advantages over the classical approach (Hamaker & Klugkist, 2011). The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) is available under the Bayesian estimation, and it can be used in similar fashion as the AIC and BIC to select the optimal model (i.e., models with small DIC values are favored). In the most recent version of Mplus 6.1 (L. K. Muthén & Muthén, 1998-2010), the Bayesian estimator is available but not yet implemented for MGMM (B. Muthén, personal communication, July 29, 2011). In addition, Bayesian analysis with a specific prior distribution allows model selection using posterior model probabilities or the Bayes factor (Hamaker & Klugkist, 2011). We expect the model selection by DIC to be similar to AIC according to

Spiegelhalter et al. (2002), whereas Bayes factor and posterior probabilities under full Bayesian analysis to be similar to the BIC, as they belong to Bayesian model selection approach. Additional research on the performance of DIC, posterior model probabilities, and Bayes factor for MGMM and GMM under the Bayesian estimation framework is needed to see how they compare with the classical/frequentist approach.

Appendix

Below is a list of all acronyms used in the current article as well as what they denote, in alphabetical order.

AIC—Akaike information criterion
 ALMR—adjusted Lo–Mendell–Rubin likelihood ratio test
 ANOVA—analysis of variance
 BIC—Bayesian information criterion
 BLRT—bootstrap likelihood ratio test
 CAIC—consistent Akaike information criterion
 FMM—factor mixture model
 GMM—growth mixture model
 IC—information criterion
 ICC—intraclass correlation
 ID—identifier
 IRT—item response theory
 LCA—latent class analysis
 LCGA—latent class growth model
 LGCM—latent growth curve model
 LR—likelihood ratio
 LRT—likelihood ratio test
 K-L—Kullback–Leibler
 MGMM—multilevel growth mixture model
 SABIC—sample size–adjusted BIC
 VLMR—Vuong–Lo–Mendell–Rubin likelihood ratio test

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Error bars computed as 1.96 standard deviations.
2. $w_i = \frac{\exp[-(1/2)\Delta_i]}{\sum_{i=1}^m \exp[-(1/2)\Delta_i]}$, m = number of models fitted,
 $\Delta_i = \text{AIC}_i - \min \text{AIC}$.

References

- Agresti, A. (1996). *An introduction to categorical data analysis*. New York, NY: John Wiley.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.
- Allua, S. S. (2007). *Evaluation of single- and multilevel factor mixture model estimation* (Order no. 3271046). ProQuest Dissertations & Theses Global (304780557). Retrieved from <https://libproxy.library.unt.edu/login?url=http://search.proquest.com/docview/304780557?accountid=7113>
- 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). Greenwich, CT: Information Age.
- Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*, 125-145. doi:10.1007/s11634-013-0154-2
- Bartolucci, F., & Murphy, T. B. (2015). A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports*, *11*, 193-203. doi:10.1515/jqas-2014-0060
- Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, *36*, 491-522. doi:10.3102/1076998610381396
- Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338-363.
- Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3-29.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. Hoboken, NJ: John Wiley.
- Boscardin, C. K., Muthén, B., Francis, D. J., & Baker, E. L. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology*, *100*, 192-208.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions. *Psychometrika*, *52*, 345-370.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261-304. doi:10.1177/0049124104268644
- Chen, Q., Kwok, O., Luo, W., & Willson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling*, *17*, 570-589.
- Clark, S., & Muthén, B. O. (2007, April). *How to handle clustered data when deciding on the number of classes in a latent class analysis: A Monte Carlo simulation study*. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.
- D'Angiulli, A., Siegel, L. S., & Maggi, S. (2004). Literacy instruction, SES, and word-reading achievement in English-language learners and children with English as a first language: A longitudinal study. *Learning Disabilities Research & Practice*, *19*, 202-213.
- Dettmers, S., Trautwein, U., Lüdtke, O., Kunter, M., & Baumert, J. (2010). Homework works if homework quality is high: Using multilevel modeling to predict the development of achievement

- in mathematics. *Journal of Educational Psychology*, *102*, 467-482. doi:10.1037/a0018453
- Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling*, *15*, 75-95.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in psychology and education* (3rd ed.). Needham Heights, MA: Allyn & Bacon.
- Graves, S., & Frohwerk, A. (2009). Multilevel modeling and school psychology: A review and practical example. *School Psychology Quarterly*, *24*, 84-94.
- Hamaker, E. L., & Klugkist, I. (2011). Bayesian estimation of multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for advanced multilevel analysis* (pp. 137-161). New York, NY: Routledge.
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling*, *14*, 202-226.
- Hox, J. (2010). *Multilevel analysis techniques and applications, second edition*. Mahwah, NJ: Lawrence Erlbaum.
- Jeffries, N. O. (2003). A note on "testing the number of mixture components in a normal mixture." *Biometrika*, *90*, 991-994. doi:10.1093/biomet/90.4.991
- Jenkins, J. M., Dunn, J., O'Connor, T. G., Rasbash, J., & Behnke, P. (2005). Change in maternal perception of sibling negativity: Within- and between-family influences. *Journal of Family Psychology*, *19*, 533-541. doi:10.1037/0893-3200.19.4.533
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics*, *62*, 49-66.
- Khoo, S., West, S. G., Wu, W., & Kwok, O. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 301-317). Washington, DC: American Psychological Association.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: SAGE.
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, *33*, 188-229. doi:10.1177/0049124103262065
- Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, *20*, 1350-1360.
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767-778.
- Lukociene, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241-249). Berlin, Germany: Springer.
- Maitra, R., & Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, *19*, 354-376. doi:10.1198/jcgs.2009.08054
- Marcus, D. K., Kashy, D. A., & Baldwin, S. A. (2009). Studying psychotherapy using the one-with-many design: The therapeutic alliance as an exemplar. *Journal of Counseling Psychology*, *56*, 537-548. doi:10.1037/a0017291
- Matsumoto, D., Nezlek, J. B., & Koopmann, B. (2007). Evidence for universality in phenomenological emotion response system coherence. *Emotion*, *7*, 57-67. doi:10.1037/1528-3542.7.1.57
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley.
- Melnykov, V., Chen, W., & Maitra, R. (2012). MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software*, *51*(12), 1-25. doi:10.18637/jss.v051.i12
- Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, *41*, 473-497.
- Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, *39*, 129-149.
- Morgan, G. B., Hodge, K. J., & Baggett, A. R. (2016). Latent profile analysis with nonnormal mixtures: A Monte Carlo examination of model selection using fit indices. *Computational Statistics & Data Analysis*, *93*, 146-161. doi:10.1016/j.csda.2015.02.019
- Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods*, *8*, 369-377.
- Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: SAGE.
- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.
- Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, *4*, 139-177.
- Ng, S. K., & McLachlan, G. J. (2014). Mixture models for clustering multilevel growth trajectories. *Computational Statistics and Data Analysis*, *71*, 43-51. doi:10.1016/j.csda.2012.12.007
- Ng, S. K., McLachlan, G. J., Wang, K., Jones, L. B., & Ng, S. W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, *22*, 1745-1752. doi:10.1093/bioinformatics/btl165
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535-569.
- Palardy, G., & Vermunt, J. K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics*, *35*, 532-565.
- Petras, H., & Masyn, K. (2010). General growth mixture analysis with antecedents and consequences of change. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 69-100). New York, NY: Springer.
- Peugh, J., & Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? A simulation study examining GMM's performance characteristics. *Structural Equation Modeling*, *19*, 204-226.
- Pruchno, R., Wilson-Genderson, M., & Cartwright, F. P. (2009). Depressive symptoms and marital satisfaction in the context of chronic disease: A longitudinal dyadic analysis. *Journal of Family Psychology*, *23*, 573-584. doi:10.1037/a0015878
- Raftery, A. E. (1996). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111-163). Oxford, UK: Basil Blackwell.

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.
- Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*, 387-401.
- SAS Institute Inc. (2002-2008). SAS, Release 9.2 [Computer program]. Cary, NC: Author.
- Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.
- Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, 64*, 583-616.
- Tobler, A. L., & Komro, K. A. (2010). Trajectories of parental monitoring and communication and effects on drug use among urban young adolescents. *Journal of Adolescent Health, 46*, 560-568. doi:10.1016/j.jadohealth.2009.12.008
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in a growth mixture models. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Greenwich, CT: Information Age.
- Vancouver, J. B. (1997). The application of HLM to the analysis of the dynamic interaction of environment, person and behavior. *Journal of Management, 23*, 795-818. doi:10.1177/014920639702300606
- Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89-106). New York, NY: Cambridge University Press.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307-333.
- Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research, 33*, 167-187. doi:10.1177/0049124104268642
- Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics & Data Analysis, 50*, 1090-1104.

Author Biographies

Qi Chen is an associate professor of Research, Measurement and Statistics within the Educational Psychology Department at UNT. Her research expertise includes growth mixture modeling, multi-level modeling, structural equation modeling, longitudinal data analysis, mediation and moderation analysis, parenting and child achievement subject matter, and school-based prevention work.

Wen Luo is an associate professor in the Department of Educational Psychology. She specializes in quantitative research methodology. Her research interests include statistical modeling of data with complex multilevel structures and longitudinal data analyses.

Gregory J. Palardy's research focuses on school and teacher effectiveness and applications of quantitative methods in educational research. Recent studies have addressed: the consequences of socioeconomic segregation in schools on cognitive outcomes, non-cognitive outcomes, and educational attainment; the impact of access to qualified and effective teachers on achievement gaps; and the impact of including summer on value-added models of teacher and school effectiveness.

Ryan Glaman is a PhD candidate in Educational Psychology with a concentration in Research, Measurement, and Statistics at the University of North Texas. His methodological research interests are in the areas of multilevel modeling and structural equation modeling

Amber McEnturff is a PhD candidate in Educational Psychology at the University of North Texas. She specializes in multilevel models, as well as data analysis and program evaluation in school and nonprofit settings.