MULTIVARIATE
BEHAVIORAL
RESEARCH
The Journal of the Society of
Multivariate Experimental Psychology

Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Bayesian Inference for Growth Mixture Models with Latent Class Dependent Missing Data

Zhenqiu Laura Lu [a], Zhiyong Zhang [a] & Gitta Lubke [a]
[a] University of Notre Dame

Available online: 08 Aug 2011

To cite this article: Zhenqiu Laura Lu, Zhiyong Zhang & Gitta Lubke (2011): Bayesian Inference for Growth Mixture Models with Latent Class Dependent Missing Data, Multivariate Behavioral Research, 46:4, 567-597

To link to this article: http://dx.doi.org/10.1080/00273171.2011.589261

Multivariate Behavioral Research, 46:567–597, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2011.589261

Psychology Press
Taylor & Francis Group

Bayesian Inference for Growth Mixture Models with Latent Class Dependent Missing Data

Zhenqiu Laura Lu, Zhiyong Zhang, and Gitta Lubke
University of Notre Dame

Growth mixture models (GMMs) with nonignorable missing data have drawn increasing attention in research communities but have not been fully studied. The goal of this article is to propose and to evaluate a Bayesian method to estimate the GMMs with latent class dependent missing data. An extended GMM is first presented in which class probabilities depend on some observed explanatory variables and data missingness depends on both the explanatory variables and a latent class variable. A full Bayesian method is then proposed to estimate the model. Through the data augmentation method, conditional posterior distributions for all model parameters and missing data are obtained. A Gibbs sampling procedure is then used to generate Markov chains of model parameters for statistical inference. The application of the model and the method is first demonstrated through the analysis of mathematical ability growth data from the National Longitudinal Survey of Youth 1997 (Bureau of Labor Statistics, U.S. Department of Labor, 1997). A simulation study considering 3 main factors (the sample size, the class probability, and the missing data mechanism) is then conducted and the results show that the proposed Bayesian estimation approach performs very well under the studied conditions. Finally, some implications of this study, including the misspecified missingness mechanism, the sample size, the sensitivity of the model, the number of latent classes, the model comparison, and the future directions of the approach, are discussed.

Longitudinal data analysis (LDA) has become widely used in medical, social, psychological, and educational research to investigate both intraindividual

Correspondence concerning this article should be addressed to Zhenqiu Laura Lu, University of Notre Dame, Department of Psychology, 118 Hagger Hall, Notre Dame, IN 46556. E-mail: Lu.30@nd.edu

changes over time and interindividual differences in changes (e.g., Demidenko, 2004; Fitzmaurice, Laird, & Ware, 2004; Hedeker & Gibbons, 2006; Singer & Willett, 2003). LDA involves data collection on the same participants through multiple wave surveys or questionnaires (e.g., Baltes & Nesselroade, 1979), so heterogeneous data are very common in practical research in these fields (e.g., McLachlan & Peel, 2000). In other words, the data collected often come from more than one distribution with different population parameters. Furthermore, during longitudinal data collection, missing data are almost inevitable because of dropout, fatigue, and other factors (e.g., Little & Rubin, 2002; Schafer, 1997).

Growth mixture models (GMMs) have been developed to provide a flexible approach to analyzing longitudinal data with mixture distributions (e.g., Bartholomew & Knott, 1999) and received a lot of attention in the literature. GMMs are combinations of *finite mixture models* (e.g., Bartholomew & Knott, 1999; Luke, 2004; McLachlan & Peel, 2000; Yung, 1997) and *latent growth curve models* (LGCs; e.g., Preacher, Wichman, MacCallum, & Briggs, 2008; Singer & Willett, 2003; Willett & Sayer, 1994). They can also be viewed as special cases of *latent variable mixture models* (Lubke & Neale, 2006) that allow patterns in the repeated measures to reflect a finite number of trajectory types, each of which corresponds to an unobserved or latent class in the population (e.g., Elliott, Gallo, Have, Bogner, & Katz, 2005; Muthén & Shedden, 1999). For a comprehensive introduction to finite mixture model theory and recent advances, see McLachlan and Peel (2000).

An important issue in the analysis of GMMs is the presence of missing data (e.g., Little & Rubin, 2002; Schafer, 1997). Little and Rubin (2002) distinguished three different missing data mechanisms: (1) *missing completely at random (MCAR)*, (2) *missing at random (MAR)*, and (3) *missing not at random (MNAR)*. MCAR is a process in which data missingness is independent of both observed and unobserved outcomes. For MAR, data missingness may depend on observed outcomes but not on unobserved outcomes. If missingness depends on unobserved outcomes or some unobserved latent variables in the fitted model, then the missingness mechanism is MNAR.

For example, in a pretest-posttest study, some students may drop out of the study after taking the pretest, and thus there are missing data due to their withdrawals. For these students, the pretest scores are observed outcomes and the posttest scores are unobserved potential outcomes. If the dropout is due to a family's move, then the missing mechanism is independent of both pretest and posttest scores; therefore it can be viewed as MCAR. If the dropout is due to a low pretest score, then the missingness depends on the pretest score but not on the posttest score and therefore it is MAR. If the dropout is due to poor performance on the posttest, then the dropout depends on the unobserved posttest score and therefore it is MNAR. If there are several latent classes in the study and the dropout is due to the latent class membership, then the dropout should also be MNAR.

The MCAR and MAR mechanisms are often referred to as *ignorable missingness* mechanisms because either the parameters that govern the missing process are distinct from the parameters that govern the model outcomes or the missingness depends on some observed variables, and therefore the likelihood-based estimates are generally consistent if the missing data mechanism is ignored (Little & Rubin, 2002).

The MNAR mechanism, on the contrary, is a *nonignorable missingness* mechanism (Little & Rubin, 2002). When the assumption of ignorable missingness mechanisms is untenable, it becomes necessary to model missingness mechanisms that contain information about the parameters of the complete data population.

Focusing on the nonignorable missingness mechanism, methods and models are available in dealing with missing data. When data come from a single population, there are two possible types of nonignorable missingness: *outcome dependant (OD)* missingness and *latent variable dependent (LVD)* missingness. OD missingness occurs when data missingness depends on the unobserved outcomes. For example, Diggle and Kenward (1994) proposed a selection model for continuous longitudinal data subject to nonignorable dropout where missingness on the current occasion is dependent on the historical observations and the current outcome that would be observed if the participant did not drop out. LVD missingness occurs when data missingness depends on some latent variables within the population, such as latent factors, latent slopes, or other latent random effects. For example, Wu and Carroll (1988) and Wu and Bailey (1989) modeled the informative right censoring process where the missingness depends on the latent rate of change. OD and LVD missingness may occur simultaneously when missingness depends on both unobserved outcomes and some latent variables. For example, Lee and Tang (2006) and Song and Lee (2007) proposed a Bayesian method for structural equation models (SEMs; e.g., Bollen, 1989; Lee, 2007) with nonignorable missingness in which the missingness may depend on the potential outcomes and the related latent variables.

When data come from mixture models, the nonignorable missingness could be OD or/and LVD missingness within mixture components and *latent class dependent (LCD)* missingness in which data missingness depends on latent random class membership. Studies that have contributed greatly to combining finite mixture models and different types of nonignorable missingness include Cai and Song (2010) and Cai, Song, and Hser (2010). Cai & Song extended Lee and Tang's (2006) single SEM with nonignorable missingness to mixture SEMs with nonignorable missingness. Cai et al. further extended the mixture SEMs to allow for missing responses in both missing outcomes and missing covariates.

The LCD missingness is an important issue in both theoretical and practical research. For example, Roy (2003) proposed a pattern mixture method to study nonignorable dropout where dropout time is related to the latent class membership. Frangakis and Rubin (1999) studied nonignorable nonresponses in a broken randomized pretest-posttest experiment by introducing a partial observed class variable, compliance, and obtained normal approximations of estimators under a series of assumptions. Using the compliance variable, Barnard, Frangakis, Hill, and Rubin (2003) studied a real data case by adopting a partial pattern mixture model to deal with missingness through Bayesian methods. Note that the LCD missingness is nonignorable because the class membership in mixture models is a latent variable, so LCD can be viewed as a special LVD missingness in mixture models.

Attrition in GMMs is very common for real data and therefore it is very important to evaluate missing data methods for GMMs. However, in the framework of GMMs, there is rare work discussing how to deal with nonignorable missingness and even less how to model the LCD missingness in GMMs. In an unpublished webnote, Muthén and Brown (2001) extended the GMMs introduced by Muthén and Shedden (1999) to deal with missing data. As a reaction to Barnard et al.'s (2003) paper, Muthén, Jo, and Brown (2003) switched from pretest-posttest models to GMMs and discussed possible approaches to bring together GMMs with missing data with latent variables.

In addition, most of previous studies rely on maximum likelihood methods for parameter estimation and carry out inferences through conventional likelihood procedures. Bayesian methods provide great advantages in the analysis of complex models with complicated data structure (e.g., Ansari, Jedidi, & Jagpal, 2000; Dunson, 2000; Scheines, Hoijtink, & Boomsma, 1999), and the application of Bayesian methods in psychological research has recently become popular through its usage by Lee and colleagues (e.g., Lee, 2007; Lee & Shi, 2000; Lee & Tang, 2006; Song & Lee, 2007; Zhu & Lee, 2001).

The goal of this article is to propose and evaluate a Bayesian approach to estimating GMMs with nonignorable missingness with a focus on LCD missingness in GMMs. Specifically, the model evaluated in this study allows (a) observed covariates to predict the class probability and (b) the latent class membership and observed covariates to predict missingness on each occasion.

This model implies that on each occasion, conditional on the class membership, the missingness given observed covariates is independent of potential outcomes. The missingness represents a form of latent ignorability (LI; Frangakis & Rubin, 1999), which states that, within each latent class, potential outcomes and associated potential response indicators are independent. LI is widely used in the analysis of broken randomized experiment for intent-to-treatment (ITT) effect and complier average causal effect (CACE; e.g., Barnard et al., 2003; Coronary Drug Project Research Group, 1980; Taylor & Zhou, 2009).

The rest of the article consists of five sections. The first describes an extended GMM where class probabilities and nonignorable missingness are modeled. The second presents the estimation of such a GMM through a full Bayesian estimation method utilizing data augmentation and Gibbs sampling algorithms. The third illustrates the application of the model and method through the analysis of mathematical ability growth data from the National Longitudinal Survey of Youth 1997 (NLSY97; Bureau of Labor Statistics, U.S. Department of Labor, 1997). The fourth presents a simulation study to evaluate the performance of the model and the Bayesian estimation method. The last section discusses the implications and future directions of this study. In addition, the Appendices present some technical details.

EXTENDED GMMS WITH LCD MISSING DATA

In this section, we present the proposed extended GMM with LCD missing data. Although focusing on the LCD missingness in this article, the model is very flexible and can be easily modified to cover a variety of missing mechanisms. The path diagram of the model is illustrated in Figure 1. In the diagram, each small square represents an observed variable, each circle represents a latent variable, a circle inside of a square represents an outcome variable with possible missing values, and the triangle represents a constant. The details of the proposed model are given as follows.

Latent Growth Curve Models (LGCs)

In Figure 1, the path diagram inside each component, the big square, illustrates an LGC model. Suppose that in a longitudinal study there are N subjects and T measurement occasions or time points. For individual i ($i = 1, 2, \ldots, N$), let $\mathbf{y}_i$ be a $T \times 1$ random vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ where y_{it} stands for the outcome or observation on occasion t ($t = 1, 2, \ldots, T$), and let $\boldsymbol{\eta}_i$ be a $q \times 1$ random vector containing q continuous latent variables. An LGC of the outcome $\mathbf{y}_i$ related to the latent $\boldsymbol{\eta}_i$ can be expressed as

$$\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \mathbf{e}_i, \tag{1}$$

where $\boldsymbol{\Lambda}$ is a $T \times q$ matrix consisting of factor loadings and $\mathbf{e}_i$ is a $T \times 1$ vector of residuals or measurement errors that are assumed to follow a multivariate normal distribution $\mathbf{e}_i \sim MN_T(\mathbf{0}, \boldsymbol{\Theta})$.[1] If we assume residual variances are

[1] Throughout the article, MN_n denotes an n-dimensional multivariate normal distribution.

○ Latent variable
□ Observed variable
▣ Observed variable with possible missing value
△ Constant

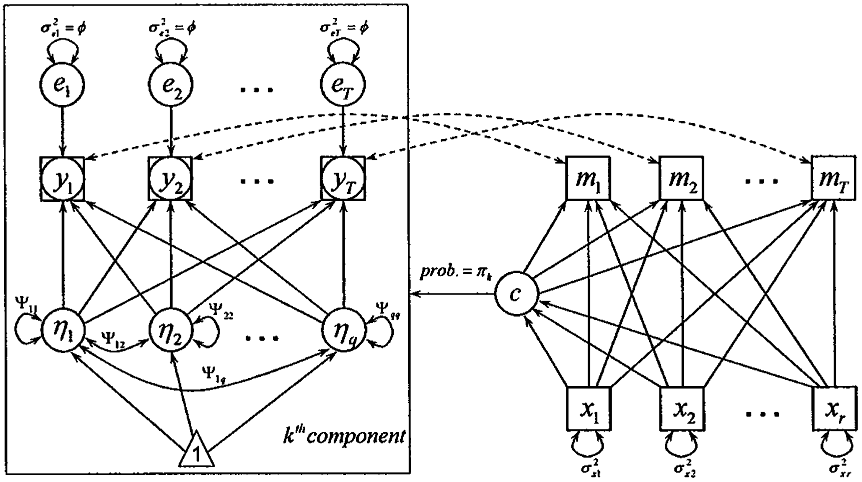

FIGURE 1 Path diagram of a growth mixture model with latent class dependent missing data. m_t indicates the missingness status of the corresponding y_t. $m_t = 1$ implies y_t is missing and $m_t = 0$ implies y_t is observed. x_rs are covariates. $p(m_t)$ depends on x_rs and the class membership c, and c is predicted by covariates x_rs. The growth mixture model takes the k^{th} component with a probability of π_k.

invariant over time, then the covariance matrix $\mathbf{\Theta} = \mathbf{I}_T \phi$, where ϕ is a scalar and $\mathbf{I}_T$ is a $T \times T$ identity matrix. The matrix $\mathbf{\Lambda}$ and the vector $\boldsymbol{\eta}_i$ determine the growth trajectory of the model. For instance, when $q = 2$, $\boldsymbol{\eta}_i = (l_i, s_i)'$, and $\mathbf{\Lambda}$ is a $T \times 2$ matrix with the first column full of 1s and the second column being $(0, 1, \ldots, (T - 1))$, the corresponding model represents a linear growth model in which l_i is the latent random level (or intercept) and s_i is the latent random slope for individual i. Furthermore, when $q = 3$, $\boldsymbol{\eta}_i = (l_i, s_i, q_i)'$, and $\mathbf{\Lambda}$ is a $T \times 3$ matrix with the first column full of 1s, the second column being $(0, 1, \ldots, (T - 1))$, and the third column being $(0, 1, \ldots, (T - 1)^2)$, the corresponding model represents a quadratic growth curve model in which l_i is the latent random level (or intercept), s_i is the latent random slope, and q_i is a latent random quadratic coefficient for individual i.

We further assume

$$\boldsymbol{\eta}_i = \boldsymbol{\beta} + \boldsymbol{\xi}_i, \tag{2}$$

where $\boldsymbol{\xi}_i$ are $q \times 1$ vectors following a multivariate normal distribution $\boldsymbol{\xi}_i \sim MN_q(\mathbf{0}, \Psi)$. $\boldsymbol{\beta}$ is called *fixed effect* and $\boldsymbol{\xi}_i$ is called *random effect* (e.g., Fitzmaurice et al., 2004; Hedges, 1994; Luke, 2004; Singer & Willett, 2003).

By combining Equation (1) and Equation (2), under the normality assumptions of both $\mathbf{e}_i$ and $\boldsymbol{\xi}_i$ and the independence assumption between $\mathbf{e}_i$ and $\boldsymbol{\xi}_i$, we have

$$\mathbf{y}_i \sim MN_T(\boldsymbol{\mu}, \Sigma),$$

where $\boldsymbol{\mu} = \boldsymbol{\Lambda}\boldsymbol{\beta}$ and $\Sigma = \boldsymbol{\Lambda}\Psi\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$.

Growth Mixture Models (GMMs)

In Figure 1, a GMM is illustrated by the LGC components and a latent categorical variable c, which stands for the latent class membership. GMMs assume that $\mathbf{y}_i$ follows a mixture of K distributions with each component distribution being a trajectory class (but see Lubke & Neale, 2008, for a discussion). The mixing proportions are also called class probabilities or weights. The density function of $\mathbf{y}_i$ is

$$p(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \, p_k(\mathbf{y}_i), \tag{3}$$

where $p_k(\mathbf{y}_i)(k = 1, \ldots, K)$ are component LGC densities, and π_k are class probabilities satisfying $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ (McLachlan & Peel, 2000).

If each mixture component $p_k(\mathbf{y}_i)$ is further assumed a multivariate normal distribution $MN_T(\boldsymbol{\mu}_k, \Sigma_k)$ where $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_k\boldsymbol{\beta}_k$ and $\Sigma_k = \boldsymbol{\Lambda}_k\Psi_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k$, then Equation (3) can be further expressed as a parametric finite normal GMM (e.g., Jordan & Xu, 1995),

$$p(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \, MN_T(\mathbf{y}_i; \boldsymbol{\beta}_k, \Psi_k, \boldsymbol{\Theta}_k, \boldsymbol{\Lambda}_k). \tag{4}$$

For different trajectory classes, $\boldsymbol{\beta}_k$, $\boldsymbol{\Lambda}_k$, Ψ_k, and $\boldsymbol{\Theta}_k$ may be different. The class-specific parameters reflect different fixed-effects and different random-effects in GMMs. For example, the overall sample can be a mixture of one subsample with low initial level and little growth and another subsample with high initial level and big growth.

Note that the class membership is unknown in mixture models. But this variable is very important to interpret mixture models. For individual i, the class membership can be expressed by a single categorical variable c_i with $c_i = k$

$(k = 1, \ldots, K)$ when $\mathbf{y}_i$ comes from the kth mixture component or class. But in later work, it is convenient to work with a K-dimensional component label vector $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})'$ in place of c_i, where z_{ik}, the kth element of $\mathbf{z}_i$, is defined to be one or zero, according to whether or not $\mathbf{y}_i$ comes from the kth class. When $c_i = k$, we have $z_{ik} = 1$ and $z_{ij} = 0$ $(\forall j \neq k)$. The vector $\mathbf{z}_i$ is distributed according to a multinomial distribution consisting of one draw from K categories with a probability π_k in the kth category,

$$\mathbf{z}_i \sim \mathit{MultiNomial}(1, \pi_1, \pi_2, ..., \pi_K). \tag{5}$$

The density function for $\mathbf{z}_i$ is $p(\mathbf{z}_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$.

Extended GMMs

Now we consider extended GMMs in which class probabilities depend on observed covariates. Notice that the GMM in Equation (4) assumes that the class probability π_k is a constant for each class, although the post hoc posterior probability can vary for each individual.[2] It is interesting to see how π_k is related to some external covariates in the mixture data analysis. For example, in addition to determining the class membership of each individual, it would be useful to see how the class membership is related to individuals' background variables such as gender and income. Note that if we include the individual variant covariates into class probability, the model is not a finite mixture anymore because the class probability is not a constant.

Let π_{ik} $(i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K)$ be the probability that individual i falls into the kth class, and let

$$\delta_{ik} = \sum_{j=1}^{k} \pi_{ij}$$

be the cumulative class probability for individual i falling into the first k classes. Note that $\delta_{iK} \equiv 1$, meaning the total class probability summing up over all K class probabilities for individual i is 1. With the definition of π_{ik} and δ_{ik}, it is easy to see that when $k = 1$, $\pi_{i1} = \delta_{i1}$; when $k = 2, 3, \ldots, \text{or}, K - 1$, $\pi_{ik} = \delta_{ik} - \delta_{i,k-1}$; and when $k = K$, $\pi_{iK} = 1 - \delta_{i,K-1}$. In this way, we order the class probabilities π_{ik} from $k = 1$ to $k = K$.

Now we build a categorical regression model (e.g., Agresti, 2002; Long, 1997) of δ_{ik} on covariates by using a probit link function[3] (e.g., McCullagh & Nelder,

[2] Here we have two probabilities that need to be distinguished. The class probability π_k is a class-specific population parameter in the model, whereas the posthoc posterior probability is an individual variable that is computed for each individual once model parameters have been estimated.

[3] Note that this is only one way to specify a regression model for categorical variables.

1989). Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ir})'$ be a $r \times 1$ vector of observed covariates that may be related to the class membership; then the probit regression[4] is built as

$$\delta_{ik} = \Phi(\varphi_{k0} + \mathbf{x}_i' \boldsymbol{\varphi}_{k1}) = \Phi[(1, \mathbf{x}_i') \cdot (\varphi_{k0}, \boldsymbol{\varphi}_{k1}')'] = \Phi(X_i' \boldsymbol{\varphi}_k), \qquad (6)$$

where the scalar φ_{k0} is an intercept, $\boldsymbol{\varphi}_{k1}$ is a $r \times 1$ vector representing coefficients for covariates $\mathbf{x}_i$, both $X_i = (1, \mathbf{x}_i')'$ and $\boldsymbol{\varphi}_k = (\varphi_{k0}, \boldsymbol{\varphi}_{k1}')'$ are $(1+r) \times 1$ vectors, and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Then the class probabilities are

$$\begin{cases} \pi_{i1} = \Phi(X_i' \boldsymbol{\varphi}_1), \\ \pi_{ik} = \Phi(X_i' \boldsymbol{\varphi}_k) - \Phi(X_i' \boldsymbol{\varphi}_{k-1}), \quad (k = 2, 3, \ldots, K-1) \\ \pi_{iK} = 1 - \Phi(X_i' \boldsymbol{\varphi}_{K-1}). \end{cases} \qquad (7)$$

For convenience, we express Equation (7) as a function $\pi_{ik} = \pi(\boldsymbol{\varphi}_k, \boldsymbol{\varphi}_{k-1}, \mathbf{x}_i)$ in the remainder of this article. As a special case, if the model has two classes, then Equation (7) is simplified as $\pi_{i1} = \Phi(X_i' \boldsymbol{\varphi}_1)$ and $\pi_{i2} = 1 - \Phi(X_i' \boldsymbol{\varphi}_1)$.

Extended GMMs With LCD Missing Data

In this subsection, we model the missingness in extended GMMs. We focus on the LCD missingness. Specifically, the missing data rate on each occasion depends on both the latent class membership $\mathbf{z}_i$ and some observed covariates $\mathbf{x}_i$. To make the model more general, we also assume that (a) the missing pattern is *intermittent*, namely, participants may return for later assessments after missing earlier assessments, and (b) the missing data rates are independent across different occasions.

Let $\mathbf{m}_i = (m_{i1}, m_{i2}, \ldots, m_{iT})'$ indicate the missingness status of $\mathbf{y}_i$. If y_{it} is missing, then $m_{it} = 1$. Otherwise, $m_{it} = 0$. Let $\tau_{it} = p(m_{it} = 1)$ be the probability that y_{it} is missing. Then, m_{it} follows a Bernoulli distribution,

$$m_{it} \sim \text{Bernoulli}(\tau_{it}). \qquad (8)$$

[4] Specifically, suppose for each k $(k = 1, 2, \ldots, K)$ there exists an underlying continuous random variable c_{ik}^*, which follows a normal distribution with mean $\varphi_{k0} + \mathbf{x}_i' \boldsymbol{\varphi}_{k1}$ and variance 1,

$$c_{ik}^* = \varphi_{k0} + \mathbf{x}_i' \boldsymbol{\varphi}_{k1} + e_i,$$

where $e_i \sim N(0, 1)$. The outcome y_i comes from the first k classes when c_{ik}^* is positive. In other words,

$$\delta_{ik} = P(c_{ik}^* > 0) = P(e_i < \varphi_{k0} + \mathbf{x}_i' \boldsymbol{\varphi}_{k1}).$$

With the class membership indicating variable $\mathbf{z}_i$, the missing probability τ_{it} can be expressed as a probit link function of $\mathbf{z}_i$ and $\mathbf{x}_i$,

$$\tau_{it} = \Phi(\mathbf{z}_i'\boldsymbol{\gamma}_{zt} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt}) = \Phi[(\mathbf{z}_i', \mathbf{x}_i') \cdot (\boldsymbol{\gamma}_{zt}', \boldsymbol{\gamma}_{xt}')'] = \Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t), \qquad (9)$$

where $\boldsymbol{\omega}_i = (\mathbf{z}_i', \mathbf{x}_i')'$ and $\boldsymbol{\gamma}_t = (\boldsymbol{\gamma}_{zt}', \boldsymbol{\gamma}_{xt}')'$ in which $\boldsymbol{\gamma}_{zt}$ is a $K \times 1$ vector $\boldsymbol{\gamma}_{zt} = (\gamma_{zt_1}, \gamma_{zt_2}, \ldots, \gamma_{zt_K})'$ and $\boldsymbol{\gamma}_{xt}$ is an $r \times 1$ vector $\boldsymbol{\gamma}_{xt} = (\gamma_{xt_1}, \gamma_{xt_2}, \ldots, \gamma_{xt_r})'$. From the distribution Equation (8) and Equation (9), we have the density function of m_{it} as a function of the class membership $\mathbf{z}_i$ and observed covariates $\mathbf{x}_i$,

$$\begin{aligned} p(m_{it}) &= [\Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)]^{m_{it}}[1 - \Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)]^{1-m_{it}} \\ &= \prod_{k=1}^{K} \left\{[\Phi(\gamma_{zt_k} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})]^{m_{it}}[1 - \Phi(\gamma_{zt_k} + \mathbf{x}_i'\boldsymbol{\gamma}_{xt})]^{1-m_{it}}\right\}^{z_{ik}}, \qquad (10) \end{aligned}$$

where $\gamma_{zt_k} = \mathbf{z}_i'\boldsymbol{\gamma}_{zt}$ for $z_{ik} = 1$ or $c_i = k$.

For convenience, in the remainder of this article the parameters $\boldsymbol{\beta}$, Ψ and ϕ are referred to as the growth curve parameters, and the parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ are referred to as the probit parameters.

BAYESIAN ESTIMATION OF THE PROPOSED MODEL

In this section, we present a full Bayesian estimation approach to the proposed extended GMMs with LCD missing data. To obtain parameter estimates through Bayesian inference, we need to calculate the probability of parameters conditionally on the data. As Bayes's theorem states that the posterior distribution of the parameters equals the product of the likelihood function of the sample data and the prior distribution of the parameters divided by the marginal distribution of the data, which is a constant and does not involve any parameter, the posterior is proportional to the likelihood times the prior.

Data Augmentation and Likelihood Function

For multidimensional models with missing data, we utilize the data augmentation method (Tanner & Wong, 1987) to obtain the likelihood function. Data augmentation refers to methods for constructing iterative optimization or sampling algorithms by introducing unobserved data or latent variables (van Dyk & Meng, 2001), and the idea of adding auxiliary variables is a useful conceptual and computational tool for many problems (Gelman, Carlin, Stern, & Rubin, 2003).

Let $\mathbf{y}_i = (\mathbf{y}_i^{obs'}, \mathbf{y}_i^{mis'})'$ where $\mathbf{y}_i^{obs}$ and $\mathbf{y}_i^{mis}$ denote observed and missing data for individual i, respectively. The direct observed-data likelihood function of $\mathbf{y}_i$ and $\mathbf{m}_i$ for the ith individual is

$$L_i(\boldsymbol{\theta}|\mathbf{y}_i, \mathbf{m}_i, \mathbf{x}_i) = \int_{\mathbf{y}_i^{mis}} \sum_{k=1}^{K} [\pi_{ik}\ p_k(\mathbf{y}_i) p(\mathbf{m}_i)]\, d\mathbf{y}_i^{mis},$$

which is very difficult to evaluate due to the high dimensional integral over an unspecified mixture structure. So data augmentation method is used by adding the auxiliary variables, the missing data $\mathbf{y}_i^{mis}$, the class membership vector $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{iK})'$, and the latent random effects $\boldsymbol{\eta}_i$, to the model. With the help of auxiliary variables, the joint likelihood function of $\mathbf{y}_i$, $\mathbf{m}_i$, $\mathbf{z}_i$, and $\boldsymbol{\eta}_i$ for the ith individual can be expressed as

$$L_i(\boldsymbol{\theta}|\mathbf{y}_i, \mathbf{m}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\eta}_i) = \prod_{k=1}^{K} \left[\pi_{ik}\ p(\mathbf{y}_i|\boldsymbol{\eta}_i)\ p_k(\boldsymbol{\eta}_i) p(\mathbf{m}_i)\right]^{z_{ik}}.$$

By combining Equations (1), (2), and (10), the likelihood function for the whole sample is

$$\begin{aligned}
L &= \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \left\{ \prod_{k=1}^{K} \left[\pi_{ik}\ p(\mathbf{y}_i|\boldsymbol{\eta}_i)\ p_k(\boldsymbol{\eta}_i) \prod_{t=1}^{T} p(m_{it}) \right]^{z_{ik}} \right\} \\
&\propto \prod_{i=1}^{N} \prod_{k=1}^{K} \{ \pi(\boldsymbol{\varphi}_k, \boldsymbol{\varphi}_{k-1}, \mathbf{x}_i) \\
&\qquad \times |\mathbf{I}_T \phi_k|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\Lambda}_k \boldsymbol{\eta}_i)' (\mathbf{I}_T \phi_k)^{-1} (\mathbf{y}_i - \boldsymbol{\Lambda}_k \boldsymbol{\eta}_i) \right] \\
&\qquad \times |\Psi_k|^{-1/2} \exp\left[-\frac{1}{2} (\boldsymbol{\eta}_i - \boldsymbol{\beta}_k)' \Psi_k^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\beta}_k) \right] \\
&\qquad \times \prod_{t=1}^{T} \left[\Phi(\gamma_{zt_k} + \mathbf{x}_i' \boldsymbol{\gamma}_{xt})\right]^{m_{it}} \left[1 - \Phi(\gamma_{zt_k} + \mathbf{x}_i' \boldsymbol{\gamma}_{xt})\right]^{1-m_{it}} \Bigg\}^{z_{ik}} \\
&\triangleq \prod_{i=1}^{N} \prod_{k=1}^{K} (v_{ik})^{z_{ik}},
\end{aligned} \tag{11}$$

where "$\triangleq$" means "is defined as."

Prior and Posterior Distributions

To use Bayesian methods, we need to specify priors for the model parameters. Lee and Song (2003) found that Bayesian estimation is not sensitive to the prior, especially for large sample size. In this study, we adopted the conjugate priors because they are commonly used in the literature of Bayesian analysis (e.g., Lee, 1981; Roeder & Wasserman, 1997; Zhu & Lee, 2001). The model parameters in this study include the growth curve parameters $\boldsymbol{\beta}_k$, Ψ_k, ϕ_k $(k = 1, 2, \ldots, K)$, and the probit parameters $\boldsymbol{\varphi}_k$ $(k = 1, 2, \ldots, K - 1)$, $\boldsymbol{\gamma}_t$ $(t = 1, 2, \ldots, T)$, so $\boldsymbol{\beta}_k$ and Ψ_k can use a multivariate normal-inverse Wishart distribution prior, ϕ_k can use an inverse Gamma distribution prior, $\boldsymbol{\varphi}_k$ can use a multivariate normal distribution prior, and $\boldsymbol{\gamma}_t$ can use a multivariate normal distribution prior. In a simpler manner, we can also directly specify the prior precision of $\boldsymbol{\beta}_k$ instead of setting it proportional to Ψ_k. Appendix A lists the details of these prior distributions.

With the likelihood function and the priors, the joint posterior distribution of the unknown parameters is readily available. However, the marginal posterior distributions (Gelman et al., 2003) of the parameters are very hard to obtain explicitly because of the requirement of high-dimensional integration. Instead, we first obtain the conditional distributions for the parameters and then utilize the Gibbs sampling method (Casella & George, 1992; Geman & Geman, 1984) to generate Markov chains for the parameters and conduct Bayesian inference.

The full conditional posterior distributions for the mixture model parameters are provided by Equation (12)–Equation (18) in Appendix B. In addition, the conditional posterior distributions for the augmented variable $\mathbf{z}_i$, the latent variable $\boldsymbol{\eta}_i$, and the missing data $\mathbf{y}_i^{mis}$ $(i = 1, 2, \ldots, N)$ are also provided by Equation (19)–Equation (21), respectively, in Appendix B.

Gibbs Sampling and Statistical Inference

With the conditional posterior distributions obtained earlier, we can generate Markov chains for the unknown model parameters by implementing a Gibbs sampling algorithm (Casella & George, 1992; Geman & Geman, 1984). The Gibbs sampling is a Markov chain Monte Carlo algorithm to obtain a sequence of samples from a joint probability distribution. Starting with a set of initial guesses of all these unknown variables, it generates instances from the conditional distribution of each variable in turn, conditionally on the current values of the other variables (Geman & Geman, 1984). The sequence of samples constructs a Markov chain that can be shown ergodic (Geman & Geman, 1984), and thus after convergence the generated value is actually from the joint distribution of all parameters. It can also be shown that each variable is also a Markov chain and converges to the marginal distribution of that variable (Robert & Casella, 2004).

Gibbs sampling is especially useful when the joint distribution is complex or unknown but the conditional distribution of each variable is available.

Specifically in our model, the unknown variables include the model parameters ϕ, Ψ, $\boldsymbol{\beta}$, $\boldsymbol{\varphi}$, $\boldsymbol{\gamma}$, the augmented variables $\mathbf{z}$, $\boldsymbol{\eta}$, and missing values $\mathbf{y}^{mis}$. The following algorithm can be used.

1. Start with a set of initial values for model parameters $\phi^{(0)}$, $\Psi^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\varphi}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, $\mathbf{z}^{(0)}$, $\boldsymbol{\eta}^{(0)}$, and $\mathbf{y}^{mis(0)}$.
2. At the sth iteration, the following parameters are generated: $\phi^{(s)}$, $\Psi^{(s)}$, $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{\varphi}^{(s)}$, $\boldsymbol{\gamma}^{(s)}$, $\mathbf{z}^{(s)}$, $\boldsymbol{\eta}^{(s)}$, and $\mathbf{y}^{mis(s)}$. To generate $\phi^{(s+1)}$, $\Psi^{(s+1)}$, $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\varphi}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$, $\mathbf{z}^{(s+1)}$, $\boldsymbol{\eta}^{(s+1)}$, and $\mathbf{y}^{mis(s+1)}$, the following procedure is implemented:
   i. Generate $\phi^{(s+1)}$ from the inverse Gamma distribution in Equation (12).
   ii. Generate $\Psi^{(s+1)}$ from the inverse Wishart distribution in Equation (13).
   iii. Generate $\boldsymbol{\beta}^{(s+1)}$ from the multivariate normal distribution in Equation (14).
   iv. Generate $\boldsymbol{\varphi}^{(s+1)}$ from the distributions in Equations (15)–(17).
   v. Generate $\boldsymbol{\gamma}^{(s+1)}$ from the distribution in Equation (18).
   vi. Generate $\mathbf{z}^{(s+1)}$ from the multinomial distribution in Equation (19).
   vii. Generate $\boldsymbol{\eta}^{(s+1)}$ from the multivariate normal distribution in Equation (20).
   viii. Generate $\mathbf{y}^{mis(s+1)}$ from the normal distribution in Equation (21).

After convergence, the statistical inference can be conducted based on the generated Markov chains. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ denote a vector of all the unknown variables in the model. The converged Markov chains can be recorded as $\boldsymbol{\theta}^{(s)}, s = 1, 2, \ldots, S$, and each parameter estimate $\hat{\theta_j}$ $(j = 1, 2, \ldots, p)$ can be calculated as $\hat{\theta_j} = \sum_{s=1}^{S} \theta_j^{(s)}/S$ with standard error (SE) obtained as the standard deviation (SD) of θ_j, $s.e.(\hat{\theta_j}) = \sqrt{\sum_{s=1}^{S}(\theta_j^{(s)} - \hat{\theta_j})^2/(S-1)}$. To get the credible (confidence) intervals, the percentiles of the Markov chains can be used.

REAL DATA ANALYSIS

In this section, we illustrate the application of the Bayesian GMM model with missing data through the analysis of mathematical ability growth data from the NLSY97 survey (Bureau of Labor Statistics, U.S. Department of Labor, 1997). Specifically, data used in the current analysis were collected yearly from 1997 to 2001 on $N = 1{,}510$ adolescents. Starting in 1997 when they were 12 years old and in the 7th grade, each adolescent was administered the Peabody

Individual Achievement Test (PIAT) Mathematics Assessment to measure their mathematical ability. The same adolescents were then measured annually till 2001 when they were 16 years old and in the 11th grade.

Table 1 shows the summary statistics for the data. Overall, the means of mathematical ability increased over time with a roughly linear trend. The missing data rates range from 4.57% to 9.47%, and the raw data show the missing pattern is intermittent. About half of the sample are male (763/1,510 = 50.5%). In order to investigate the possible number of latent classes, we draw a histogram with its smoothing density estimate for mathematical ability data at each wave. The histograms are shown in Figure 2 and clearly show the bi-modes of mathematical ability for the current sample of adolescents. Therefore, a Bayesian linear GMM with two latent classes is fitted to the data in the current analysis.

For the sake of comparison, we fit two models to the data. The first one is the Bayesian GMM model we proposed and assumes that the missing data are nonignorable, and the second one assumes that the missing data are ignorable. For the first model, we evaluate whether missingness is related to class membership and the covariate sex. Because the purpose of the current analysis is to demonstrate the application of the proposed Bayesian GMM model, we adopt the priors discussed earlier with hyperparameters chosen to carry little prior information for our model parameters (Congdon, 2003; Gill, 2002; Zhang, Hamagami, Wang, Grimm, & Nesselroade, 2007). Specifically, for $\boldsymbol{\varphi}_1$, we set $\boldsymbol{\mu}_{\varphi_1} = \mathbf{0}_2$ and $\Sigma_{\varphi_1} = 10^6\mathbf{I}_2$. For ϕ_k $(k = 1, 2)$, we set $\upsilon_{0k} = s_{0k} = 0.002$. For $\boldsymbol{\beta}_k$, it is assumed that $\boldsymbol{\beta}_{k0} = \mathbf{0}_2$ and $\Sigma_{k0} = 10^6\mathbf{I}_2$. For Ψ_k, we define $m_{k0} = 2$ and $\mathbf{V}_{k0} = \mathbf{I}_2$. Finally, for $\boldsymbol{\gamma}_t^*$, we let $\boldsymbol{\gamma}_{t0}^* = \mathbf{0}_3$ and $\mathbf{D}_{t0}^* = 10^6\mathbf{I}_3$. The starting values are then set at $\varphi_1 = 0$, $\phi_k = 1$, $\boldsymbol{\beta}_k = 1$, $\Psi_k = \mathbf{I}_2$, and $\boldsymbol{\gamma}_t^* = \mathbf{0}_3$. In both prior and starting value specifications, $\mathbf{0}_d$ and $\mathbf{I}_d$ denote a d-dimensional zero vector and a d-dimensional identity matrix, respectively. For the second model, the missingness is assumed to be ignorable and therefore there is no estimate for the missingness parameters. For other model parameters, the same priors and starting values as those in the first model are used.

TABLE 1
Summary Statistics for PIAT Math Data Set

	Grade 7	*Grade 8*	*Grade 9*	*Grade 10*	*Grade 11*
M	18.147	20.041	21.178	22.465	23.110
SD	6.219	6.526	6.601	6.435	6.643
Missing data (count)	83	69	120	115	143
Missing data (percentage)	5.497	4.570	7.947	7.616	9.470
	Male	$N = 763$		Female	$N = 747$

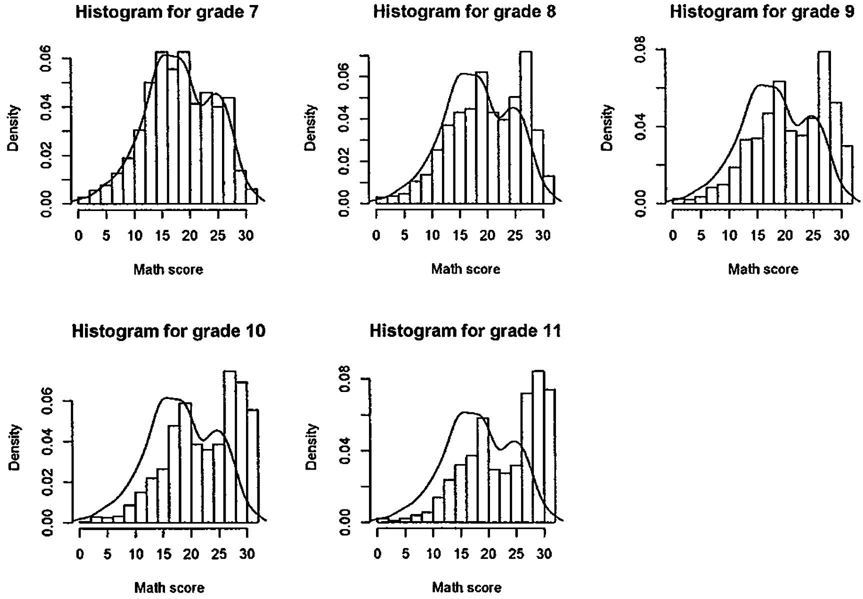

FIGURE 2 Histograms of PIAT math scores for five grades.

In generating Markov chains through Gibbs sampling, we use a burn-in period of 10, 000 iterations.[5] For testing convergence, we examine the history plot and the Geweke's z statistic (Geweke, 1992)[6] for each unknown model parameter. To make sure all the parameters are estimated accurately, the next 70, 000 iterations[7] are then saved for data analysis.

The results for our real data analysis are given in Tables 2 and 3. First, based on the history plots (two selected history plots are presented in Figure 3), it seems that each Markov chain converges to its stationary distribution. Second, the Geweke test statistics for all model parameters are smaller than 1.96, which also indicates the convergence of Markov chains (Geweke, 1992). Third, the ratio of Monte Carlo error and standard deviation for each parameter is smaller than 0.05, which indicates parameter estimates are accurate (Spiegelhalter, Thomas,

[5] With 10,000 burn-ins, the Markov chains for all parameters converged.

[6] This method tests the convergence of Markov chain by comparing the means of two subsets of the chain.

[7] With 70,000 iterations, the ratio of MCse/sd is less than 0.05 for all parameters, which indicates that the estimates are accurate. An example of inaccurate estimates obtained with 2,000 burn-ins and 5,000 iterations can be found on our website: (http://nd.psychstat.org/research/luzhanglubke2010) for comparison.

TABLE 2
Real Data Analysis Under an Assumption of Latent Class Dependent Missingness

Parameter	*M*	*SD*	*MCse*	$\frac{MCs.e.}{S.D.}$	*CI.L*[a]	*CI.U*	*Geweke t*
Growth Curve Parameters							
Class 1[b]							
$\beta_1[1]$	25.140	0.176	0.003	0.017	24.790	25.480	0.564
$\beta_1[2]$	1.130	0.039	0.001	0.026	1.055	1.206	−0.307
$\Psi_1[11]$	5.501	0.687	0.010	0.015	4.265	6.944	0.126
$\Psi_1[22]$	0.187	0.031	0.000	0.000	0.131	0.253	−0.120
$\Psi_1[12]$	−0.843	0.136	0.002	0.015	−1.127	−0.596	−0.092
ϕ_1	1.904	0.107	0.002	0.019	1.701	2.121	−0.457
Class 2[c]							
$\beta_2[1]$	15.920	0.164	0.002	0.012	15.600	16.250	1.923
$\beta_2[2]$	1.253	0.042	0.001	0.024	1.169	1.335	−0.919
$\Psi_2[11]$	15.850	1.141	0.019	0.017	13.720	18.170	1.566
$\Psi_2[22]$	0.402	0.084	0.003	0.036	0.250	0.574	−0.497
$\Psi_2[12]$	0.805	0.224	0.006	0.027	0.349	1.229	−0.476
ϕ_2	13.310	0.364	0.006	0.016	12.610	14.040	−0.750
Probit Parameters							
Class 6							
φ_{10}[d]	−0.249	0.115	0.005	0.043	−0.471	−0.024	−0.591
φ_{11}	−0.238	0.074	0.003	0.041	−0.387	−0.094	0.463
Grade 7							
γ^*_{01}[e]	−1.470	0.181	0.007	0.039	−1.835	−1.116	−0.895
γ^*_{11}	0.116	0.134	0.003	0.022	−0.135	0.397	0.344
γ_{x1}	−0.150	0.107	0.004	0.037	−0.360	0.065	1.067
Grade 8							
γ^*_{02}	−2.199	0.229	0.011	0.048	−2.662	−1.771	−1.407
γ^*_{12}	0.442	0.190	0.008	0.042	0.093	0.853	1.299
γ_{x2}	0.101	0.107	0.004	0.037	−0.113	0.309	1.146
Grade 9							
γ^*_{03}	−1.346	0.171	0.007	0.041	−1.680	−1.013	−0.835
γ^*_{13}	0.199	0.131	0.004	0.031	−0.050	0.466	1.034
γ_{x3}	−0.147	0.096	0.004	0.042	−0.333	0.038	0.655
Grade 10							
γ^*_{04}	−1.662	0.174	0.007	0.040	−2.016	−1.333	0.452
γ^*_{14}	0.192	0.131	0.004	0.031	−0.062	0.456	−0.038
γ_{x4}	0.054	0.096	0.004	0.042	−0.134	0.244	−0.577
Grade 11							
γ^*_{05}	−1.507	0.170	0.008	0.047	−1.848	−1.178	−0.854
γ^*_{15}	0.417	0.133	0.005	0.038	0.166	0.685	1.389
γ_{x5}	−0.089	0.092	0.004	0.043	−0.273	0.088	0.134

[a] The significance of parameter estimates can be judged based on the confidence intervals. If zero is included in the interval, then the parameter estimate is not significantly different from zero. [b] The growth curve parameters for Class 1. Specifically, $\beta[1]$: initial level; $\beta[2]$: slope; $\Psi[11]$: variance of initial level; $\Psi[22]$: variance of slope; $\Psi[12]$: covariance of initial level and slope; ϕ: variance of error. [c] The growth curve parameters for Class 2. [d] The probit parameters of class proportion as in Equation (6). [e] The probit parameters of missing data rate. Note that although the γ^*_{0t} and γ^*_{1t} here are different with the γ_{zt_1} and γ_{zt_2} in Equation (9), they are equivalent after reparameterizing $\gamma^*_{0t} = \gamma_{zt_1}$ and $\gamma^*_{1t} = \gamma_{zt_2} - \gamma_{zt_1}$.

Best, & Lunn, 2003). Overall, we can conclude that the results from our real data analysis can be used for further inference. For example, the distance between the two populations with different covariance matrices (Anderson & Bahadur, 1962) can be calculated and it is 2.7.

TABLE 3
Real Data Analysis Under an Assumption of Ignorable Missingness

Parameter	*M*	*SD*	*MCse*	$\frac{MCs.e.}{S.D.}$	*CI.L*	*CI.U*	*Geweke t*
Growth Curve Parameters							
Class 1							
$\beta_1[1]$	25.140	0.176	0.004	0.023	24.790	25.480	−0.239
$\beta_1[2]$	1.131	0.039	0.001	0.026	1.055	1.208	0.267
$\Psi_1[11]$	5.471	0.692	0.011	0.016	4.220	6.935	−0.684
$\Psi_1[22]$	0.187	0.031	0.000	0.000	0.130	0.252	−0.756
$\Psi_1[12]$	−0.839	0.138	0.002	0.014	−1.127	−0.586	0.927
ϕ_1	1.905	0.106	0.002	0.019	1.706	2.121	0.206
Class 2							
$\beta_2[1]$	15.890	0.164	0.002	0.012	15.570	16.210	0.760
$\beta_2[2]$	1.253	0.042	0.001	0.024	1.171	1.336	−0.578
$\Psi_2[11]$	15.640	1.116	0.018	0.016	13.550	17.920	−0.445
$\Psi_2[22]$	0.398	0.081	0.003	0.037	0.248	0.564	−0.841
$\Psi_2[12]$	0.815	0.220	0.006	0.027	0.369	1.230	0.971
ϕ_2	13.320	0.363	0.006	0.017	12.620	14.050	0.906
Class 3							
φ_{10}	−0.240	0.111	0.005	0.045	−0.460	−0.033	0.262
φ_{11}	−0.238	0.072	0.003	0.042	−0.375	−0.098	−0.249

Note. With the same notations as those in Table 2.

A quick comparison of results from both analyses shows that the estimates for growth curve parameters are very close. For both models, the differences between Class 1 and Class 2 include (a) Class 1 has a higher initial level and lower slope, (b) Class 2 has larger variations for initial level and slope, (c) the residual variance is much larger for the second class, and (d) for Class 1, the initial level and slope are negatively correlated but for Class 2 they are positively correlated.

A closer look at the results from the analysis with LCD missingness in Table 2 further reveals that none of γ_{xt}s, the coefficients for the covariate sex,

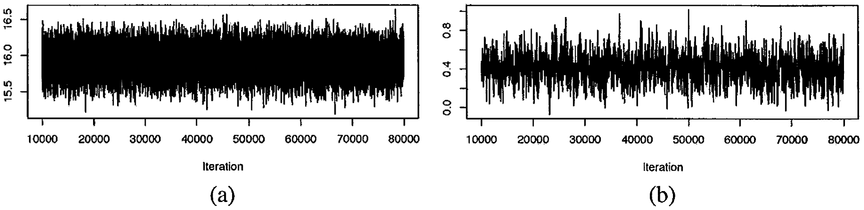

FIGURE 3 Selected history plots. History plots for all parameters can be found on our web page. (a) Parameter $\beta_2[1]$. (b) Parameter γ^*_{15}.

are significant at the α level of 0.05, which implies that the missingness may not be related to sex. However, it can be seen that the coefficients for class membership for Grades 8 and 11 are positive and significant. This indicates that Class 2 has higher missing data rates for Grades 8 and 11 than Class 1, which implies that in these two grades adolescents in Class 2 are more likely to have missing data than those in Class 1.

A SIMULATION STUDY

In this section, a simulation study is presented to evaluate the performance of the proposed Bayesian GMMs with missing data. To simplify the presentation, we focus on a linear GMM with two latent trajectory classes resembling our real data analysis. Five occasions of data are generated, and missing data are created on each occasion according to different predesigned missing data rates. It is also assumed there is only one covariate in the simulation study.

Simulation Design

In the simulation, we consider three main factors: the sample size, the class probability, and the missing data mechanism. First, the sample sizes of 1,500, 1,000, and 500 are considered. Second, both equal and unequal class probabilities are considered. For the equal class probabilities, each class contains around 50% of participants. For the unequal class probabilities, around 30% of participants are in the first class and the other 70% are in the second class. Third, both nonignorable and ignorable missing mechanisms are considered. In the simulation, we apply our model to both MCAR and MNAR data. For MCAR data, a uniform missing data rate, around 16%, is set across all five occasions for both classes. For MNAR data, the missing data rates for the first class are set around $(2\%, 4\%, 6\%, 8\%, 10\%)$ across Occasions 1 to 5, respectively, and for the second class around $(4\%, 8\%, 12\%, 16\%, 20\%)$. Different missing data rates are realized by setting different values of the corresponding probit parameters γ^*_{0t}, γ^*_{1t},[8] and γ_{xt}. The covariate x follows a normal distribution with mean 1 and standard deviation 1. In total, we evaluate the performance of the model in $3 \times 2 \times 2 = 12$ different cells.

Simulation Implementation

In the simulation, the following procedure is operated automatically.

[8] To be consistent with the real data analysis, γ_{zt_1} and γ_{zt_2} are reparameterized as γ^*_{0t} and γ^*_{1t} with $\gamma^*_{0t} = \gamma_{zt_1}$ and $\gamma^*_{1t} = \gamma_{zt_2} - \gamma_{zt_1}$.

1. Set the counter $R = 0$.
2. Generate complete GMM data according to predefined model parameters.
3. Create missing data according to missing data mechanisms and missing data rates.
4. Generate Markov chains for model parameters through the Gibbs sampling procedure.
5. Test the convergence of generated Markov chains using the Geweke statistics (Geweke, 1992).
6. If the Markov chains pass the convergence test, then set $R = R + 1$ and calculate and save the parameter estimates. Otherwise, set $R = R$ and discard the current replication of the simulation.
7. Repeat the aforementioned process till $R = 100$ to obtain 100 replications of valid simulation.

Researchers have found that there exists a label-switching problem in mixture models (e.g., Fruhwirth-Schnatter, 2001; Tueller, Drotar, & Lubke, 2011). In our analysis, we imposed some constraints on the priors to avoid the problem; for example, the intercept of the first class is constrained to be larger than that of the second class.

Because the simulation design is based on the real data analysis, the same set of uninformative priors and starting values as in the previous section (see Real Data Analysis section) are used for all simulation conditions. In generating Markov chains through the Gibbs sampling method, the burn-in period is set from 1 to $10,000$ iterations and the Markov chains with a length of $40,000$ iterations are saved for data analysis.

In this study, the Gibbs sampling algorithm is implemented in open-source software OpenBUGS (Thomas, O'Hara, Ligges, & Sturtz, 2006). OpenBUGS is flexible in estimating both simple and complex statistical modeling with a language similar to the R programming language. Lunn, Spiegelhalter, Thomas, and Best (2009), Zhang et al. (2007), and Zhang, McArdle, Wang, and Hamagami (2008) offer an overview of the use of OpenBUGS. For an in-depth account of it, see Congdon (2003) and Ntzoufras (2009). Sample OpenBUGS codes for our current models are available on our website.

Results

For the purpose of presentation, let θ_j represent the jth parameter as well as its true value in the simulation. Let $\hat{\theta}_{ij}$ denote the estimate of θ_j in the ith simulation replication. Let $\hat{s}_{ij}$ denote the estimated standard error of $\hat{\theta}_{ij}$. And let $\hat{\theta}^l_{ij}$ and $\hat{\theta}^u_{ij}$ denote the lower and upper limits of the 95% highest posterior density credible interval (HPD; Box & Tiao, 1973), respectively. For each of

12 conditions in the simulation design, we calculate five statistics defined here based on 100 sets of converged simulation replications.

First, the average estimate (Est.avg$_j$) across the 100 converged simulation replications of each parameter is obtained as Est.avg$_j = \bar{\hat{\theta}}_j = \sum_{i=1}^{100} \hat{\theta}_{ij}/100$. Second, the relative bias (Bias.rel$_j$) of each parameter is calculated using $(\bar{\hat{\theta}}_j - \theta_j)/\theta_j$ when $\theta_j \neq 0$ and $(\bar{\hat{\theta}}_j - \theta_j)$ when $\theta_j = 0$. Third, the empirical standard deviation (SD.emp$_j$) of each parameter is obtained as SD.emp$_j$ $=$ $\sqrt{\sum_{i=1}^{100} (\hat{\theta}_{ij} - \bar{\hat{\theta}}_j)^2/99}$, and fourth, the average standard deviation (SD.avg$_j$) of the same parameter is calculated by SD.avg$_j = \sum_{i=1}^{100} \hat{s}_{ij}/100$. Fifth, the coverage probability of the 0.95 HPD credible interval (HPD.cvr$_j$) of each parameter is obtained using HPD.cvr$_j = [\#(\hat{\theta}^l_{ij} \leq \theta_j$ and $\theta_j \leq \hat{\theta}^u_{ij})]/100$.

For the sake of saving space and facilitating comparison, instead of presenting full results for each condition, we further calculate four summary statistics across all model parameters for each condition of simulation. The detailed results for each condition can be found at our web page. First, we define the average absolute relative biases (|Bias.rel|) across all model parameters as $|\text{Bias.rel}| = \sum_{j=1}^{p} |\text{Bias.rel}_j|/p$. Second, we obtain the average absolute differences between the empirical *SD*s and the average Bayesian *SD*s (|SD.diff|) across all model parameters by using $|\text{SD.diff}| = \sum_{j=1}^{p} |\text{SD.emp}_j - \text{SD.avg}_j|/p$. Third, we calculate the average coverage probabilities (HPD.cvr) across all model parameters by using $\text{HPD.cvr} = \sum_{j=1}^{p} \text{HPD.cvr}_j/p$. In the aforementioned equations, p is the total number parameters in a model. These three statistics from all 12 simulation conditions are given in Table 4.

Based on the results in Table 4, we can conclude the following. First, the proposed Bayesian method can recover model parameters very well because (a) the relative biases are all small (e.g., the maximum bias is about 6.8%, which occurs when the sample size is 500, the class probability is unequal, and the missingness is MNAR) and (b) the average coverage probabilities are all close to the nominal value 95%. The correct coverage probabilities also indicate that we can use the estimated confidence intervals to conduct statistical inference. Second, with the increase of the sample size, (a) the relative biases get smaller, which shows that estimates get closer to their true values, and (b) the average Bayesian *SD*s get closer to the empirical *SD*s, which shows that standard errors become more accurate. Third, the small difference between the empirical *SD* and the average Bayesian *SD* in all conditions not only demonstrates that the Bayesian method used in the study can estimate the standard errors very well but also indicates that throwing away the nonconverged cases in our simulation does not influence the simulation results. Fourth, this model works equally well for both the MNAR missingness and the MCAR missingness. In both cases, the parameter estimate biases are small, the differences between empirical *SD*s and

TABLE 4
Summary and Comparison of Simulation Results.
The Results Are Based on the Converged Replications.

		Equal Classes		*Unequal Classes*	
		MNAR	*MCAR*	*MNAR*	*MCAR*
Sample Size					
1,500	|Bias.rel|[a]	0.022	0.009	0.026	0.011
	|SD.diff|[b]	0.011	0.008	0.011	0.008
	HPD.cvr[c]	0.956	0.941	0.949	0.954
1,000	|Bias.rel|	0.023	0.012	0.032	0.016
	|SD.diff|	0.012	0.010	0.012	0.015
	HPD.cvr	0.950	0.951	0.952	0.948
500[d]	|Bias.rel|	0.030	0.012	0.068	0.016
	|SD.diff|	0.015	0.020	0.042	0.021
	HPD.cvr	0.952	0.945	0.954	0.952

[a] The average absolute relative bias across all model parameters, defined by $|\text{Bias.rel}| = \sum_{j=1}^{p} |\text{Bias.rel}_j|/p$. [b] The average absolute difference between the empirical *SD*s and the average Bayesian *SD*s across all model parameters, defined by $|\text{SD.diff}| = \sum_{j=1}^{p} |\text{SD.emp}_j - \text{SD.avg}_j|/p$. [c] The average coverage probability across all model parameters, defined by $\text{HPD.cvr} = \sum_{j=1}^{p} \text{HPD.cvr}_j/p$. [d] With a sample size of 500, the convergence rate under unequal classes and MNAR missingness is $100/147 \approx 67\%$. MNAR = missing not at random; MCAR = missing completely at random.

average Bayesian *SD*s are tiny, and the coverage probabilities are close to the nominal level 95%.

DISCUSSION

This article presents a Bayesian method to estimate an extended GMM with LCD missingness. This model is a further extension of the finite mixture model proposed by Muthén and Shedden (1999). Instead of using the maximum likelihood estimation method, we employ a full Bayesian method. The simulation study shows that the Bayesian approach performs well, especially when the sample size is large. In the following paragraphs, we discuss six specific aspects of our study in more detail.

Misspecified Missingness Mechanism

It might be expected that mispecification of the missingness mechanism may cause a substantial misclassification of participants. For the purpose of illustration we conducted a small additional simulation. In this additional simulation,

TABLE 5
Classification Under Ignorable and Nonignorable Missingness Mechanism Assumptions

		Missingness Mechanism Assumption			
		Nonignorable[a]		*Ignorable*[b]	
		Class 1	*Class 2*	*Class 1*	*Class 2*
True Model[c]					
Class 1	506	437	69	502	4
Class 2	994	81	913	991	3
Total	1,500	518	982	1,493	7

[a] Modeling GMM and latent class dependent missingness. [b] Modeling GMM only, ignore the missingness mechanism. [c] GMM with latent class dependent missing data.

the sample size is 1,500; the class proportion is (30%, 70%); and $\beta_1 = 0$, $\Psi_{11} = \Psi_{22} = 0.5$, $\Psi_{12} = \Psi_{21} = 0$, $\phi = 1$ for both classes, $\beta_2[1] = 0$ for Class 1 and $\beta_2[2] = 1.3$ for Class 2. The distance between the two populations (Anderson & Bahadur, 1962) is 1.73. LCD missing data are then generated with different missing data proportions for different classes. The generated data are analyzed using two models. The first one uses the proposed method with missingness mechanism. For the second one, the settings are kept the same except that the missingness mechanism is ignored. To keep this article to a reasonable length, the simulation results are uploaded to our website. Table 5 gives the number of misclassified participants under ignorable and nonignorable missingness assumptions. The results clearly show that modeling nonignorable missingness as ignorable missingness can cause severe misclassification. This topic will be further investigated in our future work.

Sample Size

Generally speaking, it is difficult to provide a rule of thumb for the requirements of the sample size to distinguish between latent classes for GMM because it depends on class separation, model complexity, and other properties of the model (Lubke & Neale, 2006, 2008). It becomes more complex if we consider the missingness mechanism in addition to GMM. It is required that the outcome variables provide enough information to estimate the probit regression model parameters well. With respect to the factors examined in this study, the model and estimation method can perform very well with a sample size of 1,500, 1,000, or 500 with small missing data rates (with the lowest one around 2% in our simulation). And our experience shows that if the lowest missing data rate is relatively large (e.g., around 5%), the proposed model and estimation method can still provide useful information with a small sample size of 200.

Sensitivity of the Model

The model discussed in this study can be viewed as an example of the selection models (e.g., Heckman, 1976; Heckman & Robb, 1986; Little & Rubin, 2002) but with a more complex form. The missing mechanism is modeled explicitly by including the latent class membership as a covariate. However, our model suffers the same sensitivity problem as any other selection model. If the missingness does not depend on the class membership but some other latent or unobserved variables, our model then becomes misspecified and thus may not get valid parameter estimates. Fortunately, as we have shown, the Bayesian method can be very flexible in modeling the missing mechanism because the conditional posteriors can be obtained relatively easily through the data augmentation algorithm. Therefore, once the missing mechanism is understood, it can be modeled following the procedure outlined in this study.

Number of Latent Classes

The model and method proposed in this study is based on GMMs with a fixed number of components. For mixture models with unknown number of components, Richardson and Green (1997) used the jump Markov chain Monte Carlo (Green, 1995) in a full Bayesian analysis. Mclachlan (1987) proposed bootstrap methods (e.g., Efron & Tibshirani, 1993) to deal with problems involved in the likelihood ratios. Lee and Song (2003) employed the Bayesian factor (e.g., Berger, 1985; Kass & Raftery, 1995) and path sampling (Gelman & Meng, 1998) in Bayesian procedures of model selection for mixtures of SEMs. These techniques can be applied to our model for the determination of the number of latent classes.

Model Comparison

There are several criteria for model comparison. The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Linde, 2002) is a recently developed model comparison criterion designed for complex hierarchical models. DIC can be viewed as a Bayesian version or generalization of the Akaike's information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC, or Schwarz criterion; Schwarz, 1978). It is particularly useful in Bayesian model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. However, currently there is no exact definition for DIC in GMM with Missing Data. The problem mainly comes from at least two areas: the mixture structure and the posterior class membership. First, for mixture models or random effects models, the log-likelihood function for $p(\mathbf{y}|\theta)$ can be an observed-data log-likelihood function, a complete-data log-likelihood function, or a conditional log-likelihood function (see Celeux, Forbes, Robert, & Titterington, 2006). Second, when calculating the deviance for the final estimated parameters, it is not clear which posterior estimate of the class membership should be plugged in each individual's likelihood function. It could be a posterior mode or a posterior mean. For GMM with Missing Data, designing an effective model comparison criterion is an interesting topic for future work.

Future Directions

The models proposed in our article can be further developed in various ways. First, the missingness can be predicted by both latent random effects and the latent class membership. Also, the outcome variables, some other covariates that could explain the missingness, and any combination of these variables can be included in the model. Although such models can be in much more complex forms, the same Bayesian estimation procedure proposed in this study can be implemented. Second, in this study, a hybrid Gibbs sampling procedure is used. When the posterior does not have an explicit form, such as the probit parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$, the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) is used to generate random numbers from the posterior. And for each missing datum a Markov chain is produced, which is not very efficient for large missing data. Thus, future research must develop a more efficient way to deal with missing data. Third, as mentioned earlier, this model may be sensitive to the missing mechanism and model specification. Therefore, a study can be conducted to evaluate how the model responds to model misspecification.

ACKNOWLEDGMENTS

Dr. Gitta H. Lubke was supported by DA 018673 by NIDA. Dr. Zhiyong Zhang was partially supported by the Faculty Research Program at the University of Notre Dame. We thank Dr. Joseph L. Rodgers and the other three anonymous reviewers for their helpful comments and suggestions.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 1919,* 716–723.

Anderson, T. W., & Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *The Annals of Mathematical Statistics, 33,* 420–431.

Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science, 19,* 328–347.

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York, NY: Academic Press.

Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association, 98.*

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis: Kendall's library of statistics 7.* New York, NY: Edward Arnold.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer-Verlag.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Hoboken, NJ: John Wiley & Sons.

Bureau of Labor Statistics, U.S. Department of Labor. (1997). *National longitudinal survey of youth 1997 cohort, 1997–2003 (Rounds 1–7)* [Computer file]. Produced by the National Opinion Research Center, the University of Chicago, and distributed by the Center for Human Resource Research, The Ohio State University, Columbus, OH, 2005. Retrieved from http://www.bls.gov/nls/nlsy97.htm

Cai, J. H., & Song, X. Y. (2010). A Bayesian analysis of mixtures in structural equation models with nonignorable missing data. *British Journal of Mathematical and Statistical Psychology, 63,* 491–508.

Cai, J. H., Song, X. Y., & Hser, Y. I. (2010). A Bayesian analysis of mixture structural equation models with non-ignorable missing responses and covariates. *Statistic in Medicine, 29,* 1861–1874.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistican, 46*(3), 167–174.

Celeux, G., Forbes, F., Robert, C., & Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis, 4,* 651–674.

Congdon, P. (2003). *Applied Bayesian modelling*. New York, NY: Wiley.

Coronary Drug Project Research Group. (1980). Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine, 303,* 1038–1041.

Demidenko, E. (2004). *Mixed models: Theory and applications*. New York, NY: Wiley.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 43,* 49–93.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, B, 62,* 355–366.

Efron, B., & Tibshirani, R. (1993). An introduction to the bootstrap. New York, NY: CRC Press.

Elliott, M. R., Gallo, J. J., Have, T. R. T., Bogner, H. R., & Katz, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics, 6,* 119–143.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

Frangakis, C. E., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika, 86,* 365–379.

Fruhwirth-Schnatter, S. (2001). MCMC estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association, 96,* 194–209.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science, 13,* 163–185.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–193). Oxford, UK: Clarendon Press.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton, FL: CRC Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82,* 711–732.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, 5,* 475–492.

Heckman, J., & Robb, R. (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 63–107). New York, NY: Springer.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York, NY: Russell Sage Foundation.

Jordan, M. I., & Xu, L. (1995). Convergence results for the em approach to mixtures of experts architectures. *Neural Networks, 8,* 1409–1431.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Lee, S. Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika, 46,* 153–160.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach.* Chinchester, UK: John Wiley & Sons.

Lee, S. Y., & Shi, J. Q. (2000). Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Annals of the Institute of Statistical Mathematics, 52,* 722–736.

Lee, S. Y., & Song, X. Y. (2003). Bayesian model selection for mixtures of structural equation models with an unknown number of components. *British Journal of Mathematical and Statistical Psychology, 56,* 145–165.

Lee, S. Y., & Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika, 71,* 541–564.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley-Interscience.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499–532.

Lubke, G. H., & Neale, M. C. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research, 43,* 592–620.

Luke, D. A. (2004). *Multilevel modeling (quantitative applications in the social sciences).* Thousand Oaks, CA: Sage Publication, Inc.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine, 28,* 3049–3082.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York, NY: John Wiley & Sons.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics, 36,* 318–324.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087–1092.

Muthén, B., & Brown, C. H. (2001). *Non-ignorable missing data in a general latent variable modeling framework.* Unpublished draft.

Muthén, B., Jo, B., & Brown, C. H. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City (with comment). *Journal of the American Statistical Association, 98,* 311–314.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463–469.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley & Sons.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B, 59,* 731–792.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. New York, NY: Springer Science & Business Media Inc.

Roeder, K., & Wasserman, L. (1997). Practical bayesian density estimation using mixtures of mormals. *Journal of the American Statistical Association, 92,* 894–902.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics, 59,* 829–836.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall/CRC.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64,* 37–52.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Song, X. Y., & Lee, S. Y. (2007). Bayesian analysis of latent variable models with nonignorable missing outcomes from exponential family. *Statistics in Medicine, 26,* 681–693.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. v. d. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 64,* 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS manual (Version 1.4).* Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82,* 528–540.

Taylor, L., & Zhou, X. H. (2009, February). *Relaxing latent ignorability in the ITT analysis of randomized studies with missing data and noncompliance.* Seattle, WA: UW Biostatistics Working Paper Series.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News, 6,* 12–17.

Tueller, S., Drotar, S., & Lubke, G. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal, 18,* 110–131.

van Dyk, D. A., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics, 10,* 1–50.

Willett, J., & Sayer, A. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363–381.

Wu, M. C., & Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics, 45,* 939–955.

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics, 44,* 175–188.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62,* 297–330.

Zhang, Z., Hamagami, F., Wang, L., Grimm, K. J., & Nesselroade, J. R. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31,* 374–383.

Zhang, Z., McArdle, J. J., Wang, L., & Hamagami, F. (2008). An SAS interface for Bayesian analysis with WinBUGS. *Structural Equation Modeling, 15,* 705–728.

Zhu, H. T., & Lee, S. Y. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika, 66,* 133–152.

APPENDIX A

Prior Distributions

For ϕ_k $(k = 1, 2, \ldots, K)$, an inverse Gamma distribution is used,

$$\phi_k \sim IG(v_{0k}/2, s_{0k}/2),$$

where v_{0k} and s_{0k} are known hyperparameters. The inverse Gamma distribution has a density function

$$p(\phi_k) \propto \phi_k^{-v_{0k}/2-1} \exp(-\frac{s_{0k}}{2\phi_k}).$$

For $\boldsymbol{\beta}_k$ $(k = 1, 2, \ldots, K)$, the multivariate normal prior is used,

$$\boldsymbol{\beta}_k \sim MN_q(\boldsymbol{\beta}_{k0}, \Sigma_{k0}),$$

where the hyperparameter $\boldsymbol{\beta}_{k0}$ is a q-dimensional vector and Σ_{k0} is a $q \times q$ matrix.

For Ψ_k $(k = 1, 2, \ldots, K)$, the inverse Wishart distribution prior is used,

$$\Psi_k \sim IW(m_{k0}, \mathbf{V}_{k0}),$$

where the hyperparameter m_{k0} is a scalar and $\mathbf{V}_{k0}$ is a $q \times q$ matrix. The inverse Wishart distribution has a density function

$$p(\Psi_k) \propto |\Psi_k|^{-(m_{k0}+q+1)/2} \exp[-\frac{1}{2}\text{tr}(\mathbf{V}_{k0}\Psi_k^{-1})].$$

For $\boldsymbol{\varphi}_k$ ($k = 1, 2, \ldots, K-1$), we use an $(r+1)$-dimensional multivariate normal distribution,

$$\boldsymbol{\varphi}_k \sim MN_{(r+1)}(\boldsymbol{\mu}_{\varphi_k}, \Sigma_{\varphi_k}),$$

where $\boldsymbol{\mu}_{\varphi_k}$, an $(r+1)$-dimensional vector, and Σ_{φ_k}, an $(r+1)\times(r+1)$ matrix, are predetermined hyperparameters.

The prior for $\boldsymbol{\gamma}_t$ ($t = 1, 2, \ldots, T$) is chosen to be a multivariate normal distribution,

$$\boldsymbol{\gamma}_t \sim MN_{(K+r)}(\boldsymbol{\gamma}_{t0}, \mathbf{D}_{t0}),$$

where $\boldsymbol{\gamma}_{t0}$, a $(K+r)$-dimensional vector, and $\mathbf{D}_{t0}$, a $(K+r)\times(K+r)$ matrix, are predetermined hyperparameters.

APPENDIX B

Posterior Distributions

Let $n_k = \sum_{i=1}^{N} z_{ik}$ be the number of individuals who are in the kth class, and notate the set $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_N)$ as $\boldsymbol{\eta}$.

Conditional posterior distribution for ϕ_k, $k = 1, 2, \ldots, K$. The conditional posterior distribution for ϕ_k is an inverse gamma distribution,

$$\phi_k|\boldsymbol{\eta}, \mathbf{y}, \mathbf{z} \sim IG\left(a_{k1}/2, b_{k1}/2\right), \tag{12}$$

where

$$a_{k1} = v_{0k} + n_k T,$$

$$b_{k1} = s_{0k} + \sum_{i=1}^{N} z_{ik}(\mathbf{y}_i - \boldsymbol{\Lambda}_k\boldsymbol{\eta}_i)'(\mathbf{y}_i - \boldsymbol{\Lambda}_k\boldsymbol{\eta}_i).$$

Conditional posterior distribution for Ψ_k, $k = 1, 2, \ldots, K$. The conditional posterior distribution for Ψ_k is an inverse Wishart distribution,

$$\Psi_k|\boldsymbol{\beta}_k, \boldsymbol{\eta}, \mathbf{z} \sim IW\left(m_{k1}, \mathbf{V}_{k1}\right), \tag{13}$$

where

$$m_{k1} = m_{k0} + n_k,$$

$$\mathbf{V}_{k1} = \mathbf{V}_{k0} + \sum_{i=1}^{N} z_{ik}(\boldsymbol{\eta}_i - \boldsymbol{\beta}_k)(\boldsymbol{\eta}_i - \boldsymbol{\beta}_k)'.$$

Conditional posterior distribution for $\boldsymbol{\beta}_k$*, k = 1, 2, . . . , K.* The conditional posterior distribution for $\boldsymbol{\beta}_k$ is a multivariate normal distribution,

$$\boldsymbol{\beta}_k|\Psi_k, \boldsymbol{\eta}, \mathbf{z} \sim MN(\boldsymbol{\beta}_{k1}, \Sigma_{k1}), \tag{14}$$

where

$$\boldsymbol{\beta}_{k1} = \left(n_k\Psi_k^{-1} + \Sigma_{k0}^{-1}\right)^{-1}\left(\Psi_k^{-1}\sum_{i=1}^{N} z_{ik}\boldsymbol{\eta}_i + \Sigma_{k0}^{-1}\boldsymbol{\beta}_{k0}\right),$$

$$\Sigma_{k1} = \left(n_k\Psi_k^{-1} + \Sigma_{k0}^{-1}\right)^{-1}.$$

Conditional posterior distribution for $\boldsymbol{\varphi}_k$*, k = 1, 2, . . . , (K − 1).* When $k = 1$, the conditional posterior distribution for $\boldsymbol{\varphi}_1$ is

$$p(\boldsymbol{\varphi}_1|\boldsymbol{\varphi}_2, \mathbf{z}, X) \propto |\Sigma_{\varphi 1}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\varphi}_1 - \boldsymbol{\mu}_{\varphi_1})'\Sigma_{\varphi 1}^{-1}(\boldsymbol{\varphi}_1 - \boldsymbol{\mu}_{\varphi_1}) + \sum_{i=1}^{N}\left\{z_{i1}\log[\Phi(X_i'\boldsymbol{\varphi}_1)] + z_{i2}\log[\Phi(X_i'\boldsymbol{\varphi}_2) - \Phi(X_i'\boldsymbol{\varphi}_1)]\right\}\right]. \tag{15}$$

When $2 \le k \le K-2$, the conditional posterior distribution of $\boldsymbol{\varphi}_k$ is

$$p(\boldsymbol{\varphi}_k|\boldsymbol{\varphi}_{k-1}, \boldsymbol{\varphi}_{k+1}, \mathbf{z}, X) \propto |\Sigma_{\varphi k}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\varphi}_k - \boldsymbol{\mu}_{\varphi_k})'\Sigma_{\varphi k}^{-1}(\boldsymbol{\varphi}_k - \boldsymbol{\mu}_{\varphi_k}) + \sum_{i=1}^{N}\left\{z_{ik}\log[\Phi(X_i'\boldsymbol{\varphi}_k) - \Phi(X_i'\boldsymbol{\varphi}_{k-1})] + z_{i,k+1}\log[\Phi(X_i'\boldsymbol{\varphi}_{k+1}) - \Phi(X_i'\boldsymbol{\varphi}_k)]\right\}\right]. \tag{16}$$

Finally, when $k = K-1$, the conditional posterior distribution of $\boldsymbol{\varphi}_{K-1}$ is

$$p(\boldsymbol{\varphi}_{K-1}|\boldsymbol{\varphi}_{K-2}, \mathbf{z}, X) \propto |\Sigma_{\varphi K-1}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\varphi}_{K-1} - \boldsymbol{\mu}_{\varphi K-1})'\Sigma_{\varphi K-1}^{-1}(\boldsymbol{\varphi}_{K-1} - \boldsymbol{\mu}_{\varphi K-1}) + \sum_{i=1}^{N}\left\{z_{i,K-1}\log[\Phi(X_i'\boldsymbol{\varphi}_{K-1}) - \Phi(X_i'\boldsymbol{\varphi}_{K-2})] + z_{iK}\log[1 - \Phi(X_i'\boldsymbol{\varphi}_{K-1})]\right\}\right]. \tag{17}$$

The $\Phi(X_i'\boldsymbol{\varphi}_k)$ in Equations (15), (), and (17) is defined by Equation (6).

Conditional posterior distribution for $\boldsymbol{\gamma}_t$, *$t = 1, 2, \ldots, T$.* The conditional posterior distribution for $\boldsymbol{\gamma}_t$ is

$$p(\boldsymbol{\gamma}_t|\mathbf{z},\mathbf{x},\mathbf{m}) \propto \exp\Bigg[-\frac{1}{2}(\boldsymbol{\gamma}_t-\boldsymbol{\gamma}_{t0})'\mathbf{D}_{t0}^{-1}(\boldsymbol{\gamma}_t-\boldsymbol{\gamma}_{t0}) + \sum_{i=1}^{N}\left\{m_{it}\log\Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)+(1-m_{it})\log[1-\Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)]\right\}\Bigg], \tag{18}$$

where $\boldsymbol{\omega}_i = (\mathbf{z}_i', \mathbf{x}_i')'$ and $\Phi(\boldsymbol{\omega}_i'\boldsymbol{\gamma}_t)$ is defined by Equation (9).

Conditional posterior distribution for $\mathbf{z}_i$, *$i = 1, 2, \ldots, N$.* The conditional posterior distribution for $\mathbf{z}_i$ is a multinomial distribution,

$$\mathbf{z}_i|\phi, \Psi, \boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\varphi}, \boldsymbol{\eta}, \mathbf{y}, \mathbf{x}, \mathbf{m} \sim Mnomial(1, \pi_{i1}^*, \pi_{i2}^*, \ldots, \pi_{iK}^*), \tag{19}$$

where $\pi_{ik}^* = v_{ik}/\sum_{i=1}^{K} v_{ik}$ with v_{ik} defined in Equation (11).

Conditional posterior distribution for $\boldsymbol{\eta}_i$, *$i = 1, 2, \ldots, N$.* The conditional posterior distribution for $\boldsymbol{\eta}_i$ is a multivariate normal distribution,

$$\boldsymbol{\eta}_i|\boldsymbol{\phi}, \Psi, \boldsymbol{\beta}, \mathbf{z}_i, \mathbf{y}_i \sim MN(\boldsymbol{\mu}_{\eta i}, \Sigma_{\eta i}), \tag{20}$$

where

$$\boldsymbol{\mu}_{\eta i} = \sum_{k=1}^{K} z_{ik}\left[\left(\frac{1}{\phi_k}\boldsymbol{\Lambda}_k'\boldsymbol{\Lambda}_k + \Psi_k^{-1}\right)^{-1}\left(\frac{1}{\phi_k}\boldsymbol{\Lambda}_k'\mathbf{y}_i + \Psi_k^{-1}\boldsymbol{\beta}_k\right)\right],$$

$$\Sigma_{\eta i} = \sum_{k=1}^{K} z_{ik}\left(\frac{1}{\phi_k}\boldsymbol{\Lambda}_k'\boldsymbol{\Lambda}_k + \Psi_k^{-1}\right)^{-1}.$$

Conditional posterior distribution for missing data $\mathbf{y}_i^{mis}$, *$i = 1, 2, \ldots, N$.* The conditional posterior distribution for the missing data $\mathbf{y}_i^{mis}$ is a normal distribution,

$$\mathbf{y}_i^{mis}|\mathbf{z}_i, \boldsymbol{\eta}_i, \boldsymbol{\phi} \sim MN\left[\sum_{k=1}^{K} z_{ik}(\boldsymbol{\Lambda}_k\boldsymbol{\eta}_i), \sum_{k=1}^{K} z_{ik}(\mathbf{I}_T\phi_k)\right], \tag{21}$$

and its dimension and location depend on the corresponding $\mathbf{m}_i$ value.