Abstract

Title: INVESTIGATING METHODS OF INCORPORATING COVARIATES IN GROWTH MIXING MODELING: A SIMULATION STUDY

Ming Li, Doctor of Philosophy, 2015

Directed By: Dr. Jeffrey R. Harring
Measurement, Statistics and Evaluation
Department of Human Development and Quantitative Methodology

The current research aims to evaluate the performance of various approaches for estimating covariates within the latent class membership regression model in the context of growth mixture models. Researchers have been searching for more efficient and accurate estimation methods for incorporating covariate information in mixture modeling in order to clearly differentiate between subjects from different groups and to make interpretation of the growth trajectories more meaningful. However, few studies have considered more complicated models such as growth mixture models where the latent class variable is more difficult to identify. To this end, two Monte Carlo simulations were conducted. In Simulation I, four estimation approaches were investigated to examine parameter recovery, variance and standard error efficacy related to both categorical and continuous covariates that defined the regression model for the latent class membership part of the model. Data generated for Simulation II include three covariates, with one dichotomous variable linked to latent class membership and the other two (one dichotomous and one continuous) associated with measurement part of the growth mixture model. Three estimation approaches were then compared using the population data generation model as well as a misspecified model.

INVESTIGATING METHODS OF INCORPORATING COVARIATES IN GROWTH MIXTURE MODELING: A SIMULATION STUDY

By

Ming Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:

Dr. Jeffrey R. Harring, Chair
Dr. Gregory R. Hancock
Dr. Laura M. Stapleton
Dr. Ji Seung Yang
Dr. Robert Croninger

Dedication

I dedicate this dissertation work to my grandma for her unconditional love and for instilling the importance of hard work and perseverance.

Acknowledgements

First I would like to thank Dr. Jeffery Harring, my academic advisor and dissertation chair, for your patient guidance, spiritual support and enthusiastic encouragement throughout my dissertation work. Without you, I would not be able to complete my doctoral study in a timely manner.

I also would like to thank my committee members who were understanding and supportive in their own ways. A special thanks to Dr. Ji Seung Yang for discussing with me about data generation for my dissertation. Thank you, Dr. Robert Croninger, Dr. Gregory R. Hancock, Dr. Laura Stapleton, and Dr. Ji Seung Yang for agreeing to serve on my dissertation committee.

Finally, I would like to express my sincere gratitude to all EDMS faculty and my graduate colleagues. Thank you for accompanying me and providing me support through all these years. Experiences at EDMS will be great assets for my future life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Growth mixture modeling (GMM; Muthén, 2001, 2004; Muthén & Muthén, 2000; Muthén & Shedden, 1999) continues to be a popular platform for practitioners in the social and behavioral sciences to examine population heterogeneity in growth characteristics of individuals' longitudinal profiles. A primary goal of GMM is to identify two or more latent classes that represent subgroups thought to manifest qualitatively distinct patterns of change over time. Despite the increase of applied studies using GMM in the literature (see, e.g., Colder, Campbell, Ruel, Richardson, & Flay, 2002; Colder, Mehta, Balanda, Campbell, Mayhew, Stanton, Pentz, & Flay, 2001; Ellickson, Martino, & Collins, 2004; Heybroek, 2011; Huang, Murphy, & Hser, 2012; Pinquart, & Schindler, 2007), there remain unanswered methodological questions concerning the use of GMMs regarding correct model specification, optimal number of latent classes and accuracy of the classification of individuals into groups (Muthén, 2004; Nagin, 1999; Petras & Masyn, 2010).

One line of research recommends that the conventional growth mixture model be extended to incorporate covariates in the mixture analysis (Muthén, 2003; 2004). Previous simulation studies and empirical research also have demonstrated that incorporating potentially important covariates that are related to the latent mixture variable may improve parameter estimates (see, e.g., Huang, Brecht, Hara, & Hser, 2010; Li & Hser, 2011; Lubke & Muthén, 2007). Petras and Masyn (2010) discussed in detail the importance of including auxiliary information in terms of antecedents (predictors and covariates) and distal outcomes of trajectory group membership in the general GMM analysis. By including auxiliary information, the conventional GMM can be extended to estimate varying class membership probability as a function of a set of covariates (i.e., for each class the values of the latent growth parameters are allowed to be influenced by covariates) and to incorporate outcomes of the latent variables. In this way, the posterior probabilities of group membership can determine the ability of the model to clearly differentiate between subjects. Also, covariates or predictors make interpretation of the growth trajectories more meaningful because of the inclusion of individual background information, and this might be a most important reason for applied researchers to include individual specific information into the growth mixture analysis. For example, in an applied study by Pinquart and Schindler (2007) changes in life satisfaction in 1,456 German retirees were investigated using the latent growth mixture modeling. One of the goals of the study was to test whether groups showing different trajectories would vary by personal characteristics such as retirement age, gender, socioeconomic status (SES), marital status, health, employment before retirement, and region, etc. Three patterns of change in life satisfaction were identified: in Group 1, satisfaction declined at retirement but remained on a stable or increasing pattern thereafter; in Group 2 satisfaction greatly increased at retirement but overall was declining; and in Group 3, satisfaction slightly increased temporarily at retirement. It was found that the three latent groups differed by most of the covariates considered in the study. For example, members of Group 1 were older when they retired and were more likely to be female and to report worse physical health. Members in Group 2 were typically younger when they retired and were more likely to be men, to be individuals of lower SES, to be unmarried, to report worse physical health, to be unemployed before retirement, and to live in the Eastern part of

Germany, and the majority of older adults in Group 3 showed a very small temporary increase in life satisfaction after retirement (Pinquart & Schindler, 2007).

Though there are numerous advantages of including auxiliary variables in GMM analysis, the choice of an approach to estimating the model has been challenging, especially considering the fact that most of the research on estimation methods have been conducted on simple latent mixture models. For example, a conventional or standard approach to including covariates in a GMM analysis may involve the following three steps: (1) the unconditional GMM (e.g., a growth mixture model without any covariates and/or distal outcomes) is fitted based only on latent class indicators to determine the number of distinct trajectory groups; (2) class membership is assigned to each individual based on their highest posterior probability of belonging to a particular class; and (3) the relation between the assigned latent class membership and subject-specific background characteristics is investigated using either mean comparison tests (e.g., *t*-tests, ANOVAs, or chi-square tests) or multinomial logistic regression models. Whether using mean comparison tests or generalized linear regression models, one issue that arises is that class membership is treated as an exact, observed variable without taking into account the error associated with estimating these probabilities (Clark & Muthén, 2009). That is, the chances of an individual being mistakenly assigned to a particular class were not considered at all, which will lead to underestimated associations between covariates and class membership (Bolck, Croon & Hagenaars, 2004) and thus should not be used in model estimation (Nagin, 2005).

Rather than treating auxiliary information as outcomes in post-hoc comparisons as is done in the conventional approach, a one-step maximum likelihood (ML) approach (see, e.g., Bandeen-Roche, Miglioretti, Zeger & Rathouz, 1997; Dayton & Macready, 1998; Van der Heijden, Dessens & Böckenholt, 1996) was recommended, which incorporates these additional concomitant variables as part of a single model. Estimation of the model proceeded permitting for the simultaneous examination of the covariates impact on the estimation of developmental trajectories and their association with the distal outcome (Huang et al., 2010; Muthén, 2004; Nagin, 2005; Roeder, Lynch & Nagin, 1999). In the one-step approach, the latent class model and the regression model are combined into one joint model, which circumvents the problem of treating most likely class membership as an exact, observed variable. This is accomplished by taking into account the error associated with the posterior probability estimates and allowing individuals to be fractional members of all classes (Clark & Muthén, 2009). However, one major issue with this method may come from the impact of either the covariate variables or the distal outcome variable on the forming of the latent classes. That is, the latent classes formed from the joint model may differ in meaning from the latent classes obtained using the indicator variables alone and thus may potentially change their substantive interpretation. Another concern, according to Vermunt (2010), is that simultaneously building the classification model and the prediction model may not fit with the logic of most applied researchers, who often work sequentially from first building the classification model then adding covariates at a secondary stage of the analysis. Other disadvantages of the one-step approach are discussed in detail by Vermunt (2010).

To independently evaluate the relation between the latent class variable and the auxiliary variables without using assigned class membership, other approaches have been developed, such as using pseudo class (PC) draws (see, e.g., Clark & Muthén, 2009; Wang, Brown & Bandeen-Roche, 2005); and the BCH approach proposed by Bolck et al. (2004). With the PC method, for the latent class analysis, multiple random samples are drawn from a multinomial distribution of posterior probabilities (for each individual) being in each class (assuming there are more than two classes) so that each individual is given a chance to fall into neighboring classes (Clark & Muthén, 2009). Asparouhov and Muthén (2013) described the PC approach in an analogous fashion to the idea behind multiple imputation in missing data analysis which makes sense in that the latent classes are considered missing. Finally the class specific information associated with the auxiliary variable(s) is obtained using the multiple imputation techniques developed by Rubin (1987).

Vermunt (2010) proposed a new three-step maximum likelihood (ML) procedure as an extension of the BCH approach based on the work of Bolck et al. (2004). With the new three-step ML approach, Vermunt used individual observations instead of a table of frequency counts to remove the limitation of using only categorical covariates, which then not only makes it possible to use both continuous and categorical predictors but makes the model estimation much more efficient (Vermunt, 2010). In this new approach, the latent class model was estimated first. Next, the most likely class variable was set based on the highest posterior probability from the latent class posterior distribution derived from the latent class analysis. With this approach, the classification uncertainty rate and the measurement error were computed to demonstrate that the most likely class variable could be treated as an imperfect measurement of latent class analysis. Thus, in the third step, the measurement error in the most likely class was taken into account. Also, auxiliary information was included in this final stage of model estimation.

This study will investigate four estimation approaches, namely, the conventional three-step approach, the one-step ML approach, the PC approach, and the three-step ML approach, by examining the association between covariates and the latent class variable under the GMM framework. Since one of the manipulated covariates is continuous, the BCH approach is not to be included in the current study because, as mentioned before, one limitation with this approach is that it can be used only for categorical covariates.

1.1 Limitations of Previous Work

Since problems with using the conventional approach were recognized (see, e.g., Clogg, 1995; Hagenaars, 1993; Roeder et al., 1999), researchers have been searching for more efficient and accurate estimation methods when incorporating auxiliary information in mixture modeling. For example, Clark and Muthén (2009) explored how different regression methods of relating latent class analysis results to auxiliary variables can impact estimation of auxiliary effects. Results showed that the one-step approach outperformed the conventional approach and the PC method in terms of recovering the true effect of the auxiliary variable on class membership. The PC method worked well when class separation was large. Vermunt (2010) compared the conventional three-step procedure, the one-step approach, the BCH approach, and his proposed three-step ML approach with respect to bias in the estimates of the covariate effects and bias in the standard error estimates when covariates were included in latent class modeling. Results showed that the BCH method and the three-step ML method demonstrated good parameter estimates and standard errors except when the classes were poorly separated. It was also found that the three-step ML method was much more efficient than the BCH method in terms of the standard deviation of parameter estimates, and it was almost as efficient as the one-step estimation approach. One limitation with these studies is that only simple latent class models for discrete responses were used. None of these studies considered more complicated models such as growth mixture models where the latent class variable is more difficult to identify. Also, although Vermunt (2010) included three categorical predictor variables in their simulation study, Clark and Muthén (2009) only considered the impact of one continuous covariate in their study. It is quite possible that in real data analytic situations many covariates of different types should be considered simultaneously when investigating parameter recovery, model estimation, and standard error accuracy.

In a recent white paper by Asparouhov and Muthén (2013), the relation between a latent class variable and an auxiliary variable in mixture modeling was examined using different approaches under different manipulated simulation design conditions. Results showed that the new three-step ML approach uniformly outperformed the PC approach for analyzing the relation between a latent class variable and an auxiliary variable independently of the latent class model estimation. Also, if the class separation was adequate the three-step ML approach had the same efficiency as the one-step approach. One major difference between this study and the other studies was that in addition to looking at the simple latent class models, more complicated models such as a growth mixture model was included to evaluate the performance of the various estimation approaches. However, in spite of the added model complexity, limitations were noted. First, only one covariate was included, which, as was mentioned above is not common in analytic situations found in practice. Second, the impact of covariate effect size (i.e., the strength of the association of the covariate(s) with the latent class membership) on the proposed new three-step estimation method was not fully investigated.

1.2 The Current Study

As was discussed earlier, incorporating covariates related to the latent class analysis may improve the ability of the mixture model to clearly differentiate between subjects because the posterior probabilities of group membership are estimated as a function of a set of covariates. On the other hand, covariates make interpretation of the growth trajectories more meaningful because of the inclusion of individual background information. Since nearly every application in longitudinal studies incorporates some covariate information and applied researchers want to know how covariates help explain group membership, it is important that the estimation of the relation between covariates and the latent class membership is accurate. Biased covariate effect estimates from either misclassification of cases and/or from using a particular algorithm will ultimately affect the results of the analysis and make the interpretation unreliable.

Therefore, this study aims to evaluate four approaches for the estimation of parameters from growth mixture models with covariate(s): (1) the conventional approach, (2) the one-step ML approach, (3) the PC approach, and (4) the three-step ML approaches. Specifically, the estimated relations between a latent class variable and covariate(s) from using the four estimation approaches will be compared. Covariates with differing effect sizes will be one major manipulated factor in the current study.

Chapter 2: Literature Review

2.1 Latent Growth Modeling

Methods for longitudinal data analysis have experienced unprecedented development since the 1990s when models for mean change such as ANOVA and MANOVA were no longer favored (Bauer, 2007) in lieu of approaches that allowed investigation of change in individuals over time. A class of useful methods to study has emerged over the past twenty years from the area of structural equation modeling (SEM), and falls under the general heading of *latent growth modeling* (McArdle, 1988; Meredith & Tisak, 1990). Latent growth models (LGMs) allow the change process to be characterized by a mathematical function common to all subjects, but whose parameterization is permitted to vary among individuals (Bollen & Curran, 2006). That is, the relative standing of an individual at a specific time point could be modeled as a function of an underlying process which has parameter values that vary randomly across individuals (Meredith & Tisak, 1990). The analytic goals in using LGMs are to understand (1) the typical behavior of the underlying change process of the phenomenon captured by the parameters of the model, (2) the extent to which these parameters, and hence phenomenon, vary across individuals, and (3) whether some of this variability can be explained by individual-specific characteristics (Hancock, Harring & Lawrence, 2013; Harring, 2009).

In one of its simplest forms, a linear function with a subject-specific intercept and slope can be specified for each individual's continuous repeated measures that demonstrate straight line change patterns. An unconditional latent linear growth model, for repeated measurements of a continuous dependent variable, can be presented by using a general SEM notation:

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

where $\mathbf{y}_i$ is a $p \times 1$ vector of repeated measures for individual i, $\boldsymbol{\eta}_i$ is a $q \times 1$ vector of individual-

specific growth factors (i.e., intercept and slope), and $\mathbf{\Lambda}_i$ is a $p \times q$ matrix of factor loadings. Assuming (for simplicity) that the outcome variable is measured at four equal-interval time points, then the factor loading matrix might be specified:

$$\mathbf{\Lambda}_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}'. \tag{2}$$

The $p \times 1$ vector of time-specific errors, $\boldsymbol{\varepsilon}_i$, captures the deviations from the data to the fitted model for each individual. These errors are assumed to be normally distributed, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{\Theta}_i)$, with mean vector of **0** and covariance matrix $\mathbf{\Theta}_i$. At the population level, individual-specific growth factors are formulated initially as the sum of fixed and random effects,

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\zeta}_i, \tag{3}$$

where $\boldsymbol{\alpha}$ is a $q \times 1$ vector of growth factor means (i.e., intercept and slope factor means), and $\boldsymbol{\zeta}_i$ is a $q \times 1$ vector of random effects reflecting individual differences of an individual's growth factors from these means. The random effects are also assumed to be normally distributed,

$\boldsymbol{\zeta}_i \sim N(\mathbf{0}, \mathbf{\Psi})$, also with mean vector of **0** and covariance matrix, $\mathbf{\Psi}$, and are uncorrelated with the individual level errors (i.e., cov($\boldsymbol{\zeta}_i, \boldsymbol{\varepsilon}_i'$) = **0**). Therefore, the key assumptions underlying the LGM are: (a) the growth patterns for all cases are from the same functional form; and (b) the repeated measures are multivariate normally distributed, which implies that the individual growth parameters and time-specific residuals are also multivariate normal (Muthén, 2004).

The linear LGM is often a starting point in many analyses, not necessarily because it is the most realistic representation of growth for modeling a particular variable, but rather because it often fits well for many processes in a fairly restricted span of the development. Of course, other types of functional forms within the LGM framework (i.e., quadratic, logarithmic, exponential – see, e.g., Choi, Harring, & Hancock, 2009; Grimm & Ram, 2009; Petras & Masyn, 2010) are possible, especially when theory dictates a more complex growth pattern or there is sufficient empirical evidence to support such elaborations.

2.1.1 Latent growth models with covariates

In many situations, it is common at some later point of an analysis to include individual attributes at the population level model to better understand determinants which explain differences in the individual trajectories. Brown (2003) also addressed that under the primary assumption of LGMs that the growth trajectory representing change in the dependent indicator variables is modeled as a single population distribution, any nonrandom deviation from the underlying population distribution must be modeled explicitly by covariates included in the study design (e.g., social economic status, age, gender). Assuming the same linear developmental pattern discussed in Equation 1 and Equation 2 above, the conditional LGM with static covariates can be used to determine which variables influence the intercepts or slopes. Time-invariant covariates, a $r \times 1$ vector, $\mathbf{x}_i$, enter the LGM at the population level through

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \tag{4}$$

where $\boldsymbol{\Gamma}$ is a $q \times r$ matrix of coefficients relating each of the covariates to the growth factors. In a similar manner as ordinary least squares regression models, covariates can be continuous or categorical, and it is assumed that the effects of covariates or predictors on the growth factors are the same for all individuals (Petras & Masyn, 2010). In the same vein, the inclusion of covariates changes the population intercept and slopes as they are now interpreted conditionally on the covariates. It is assumed that the residuals are uncorrelated with the covariates.

2.2 Growth Mixture Models

Although the conventional LGM has the advantage of analyzing longitudinal data from the perspective of individual growth patterns, the model assumes that observed data are sampled from one homogenous population (Wang & Bodner, 2007). That is, the LGM assumes that there is a common growth pattern or trajectory for all the individuals in the study (Tofighi & Enders, 2008). In other words, the population being studied is homogeneous in terms of their growth trajectories. The fact is, however, that the observed data might reveal different subpopulations and each subpopulation has its own growth patterns defined by a particular set of model parameters. Thus, if the data being studied indicate the existence of subpopulations, the use of only one common model to describe growth for the subpopulations would not be appropriate. In this situation, analytic methods that are capable of allowing for, and actually identifying, the developmental trajectories of subpopulations are needed (Liu, 2012).

In response to the need to discern latent trajectory subgroups, a modeling technique known as growth mixture modeling (GMM; Muthén, 2001, 2004; Muthén & Muthén, 2000; Muthén & Shedden, 1999) was developed. Conceptually, GMM is a combination of latent growth modeling and latent class analysis (LCA; McCutcheon, 1987). The combination of these two methods makes it possible to identify and estimate the subpopulations with qualitatively distinct patterns of development over time (Wang & Bodner, 2007). GMM permits heterogeneity in the growth trajectories represented by a latent categorical variable that defines *k* latent classes of individuals (Tofighi & Enders, 2008). With the integration of categorical latent variables, GMM relaxes the single population assumption to allow for parameter differences across unobserved subpopulations, which means that different classes of individuals are allowed to vary around class-specific mean growth curves (Muthén, 2004). This is different from LGMs where individuals vary around a single mean growth curve. Therefore, with GMM, not only can each class have a unique set of parameters that describe its growth pattern, but within-individual and between-individual variability can also be class-specific (Wang & Bodner, 2007). This modeling flexibility is the basis of GMM framework (Muthén & Asparouhov, 2009). The linear LGM discussed above can be extended to the GMM formulation in the following way. Suppose a linear LGM is specified for each subpopulation *k*. Then the GMM model takes the following form:

$$\mathbf{y}_i^k = \boldsymbol{\Lambda}_i \boldsymbol{\eta}_i^k + \boldsymbol{\varepsilon}_i^k \tag{5}$$

$$\boldsymbol{\eta}_i^k = \boldsymbol{\alpha}^k + \boldsymbol{\zeta}_i^k , \tag{6}$$

where

$$\boldsymbol{\varepsilon}_i^k \sim N(\mathbf{0}, \boldsymbol{\Theta}_i^k) \qquad \boldsymbol{\zeta}_i^k \sim N(\mathbf{0}, \boldsymbol{\Psi}^k), \qquad k = 1, ..., K.$$

Here, differences in $\boldsymbol{\alpha}^k$ capture the differences in the growth factor means of the latent classes. The growth factor variances and covariances are also class specific, having covariance matrix $\boldsymbol{\Psi}^k$, that follows a normal distribution centered at a mean vector of 0. It is assumed that the residuals are normally distributed, have mean vector of 0 and covariance matrix $\boldsymbol{\Theta}_i^k$ which captures differences in the dispersion of the individual trajectories and time-specific residuals within classes (Bauer, 2007). It is also assumed that $\boldsymbol{\varepsilon}_i^k$ and $\boldsymbol{\zeta}_i^k$ are independent.

When estimating GMM parameters, there are additional parameters compared to the latent growth model, namely class-specific proportions φ^k of latent classes $k = 1, 2, ..., K$. Let $p(\mathbf{y}_i)$ denote the unconditional (or marginal) probability of observing individual *i*'s longitudinal sequence of measurements $\mathbf{y}_i$, and $p(\mathbf{y}_i \mid C_i = k)$ is the conditional probability distribution of $\mathbf{y}_i$ given membership in class *k*. So by aggregating the *K* conditional probability distribution functions $p(\mathbf{y}_i \mid C_i = k)$, the probability distribution of the data $\mathbf{y}_i$ is a weighted sum of the component probability distributions:

$$p(\mathbf{y}_i) = \sum_{k=1}^{K} \varphi^k p(\mathbf{y_i} \mid C_i = k), \tag{7}$$

where the latent class probabilities φ^k are constrained to be $0 < \varphi^k < 1$ and must sum to 1: $\sum_{k=1}^{K} \varphi^k = 1$. This is the sum across all *K* classes of the probability of $\mathbf{y}_i$ given subject *i*'s membership in class *k* weighted by the probability of membership in class *k*. Rolfe (2010) showed that the likelihood of the sample of *n* subjects is the product of the individual contributions to the likelihood function specified by Equation 7, namely, $L = \prod_{i=1}^{n} p(\mathbf{y}_i)$.

2.2.1 Growth mixture models with covariates

Although growth mixture models have the advantage of enumerating possible subpopulations, one challenging issue has been the identification of the “correct” number of latent classes. Class enumeration has received a great deal of attention in the methodological literature and has been investigated from various perspectives. For example, several studies compared model fit measures and statistical tests used to guide class enumeration (see, e.g., Liu & Hancock, 2014; Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2008; Wang & Bodner, 2007). Specifically, the performance of information-based indices and nested model likelihood ratio tests for relative model comparisons were studied. However, though a variety of suggestions were provided from these studies, there is no agreement on the best criteria for determining the number of classes in mixture modeling (Nylund et al., 2007). In a recent simulation study, Liu and Hancock (2014) proposed the idea of using an unrestricted multivariate normal mixture strategy to assess class enumeration. They compared the performance of a linear GMM against that of a completely unrestricted multivariate normal mixture model in terms of their ability to identify the correct number of latent classes and found that the theoretically compelling completely unrestricted multivariate normal mixture model was superior to the linear GMM when the nature of the growth curve was not certain and the sample size was sufficiently large. In addition to model comparisons and modeling strategies, another line of research has taken into account the inclusion of covariates in GMM. According to Bauer and Curran (2003), the common practice of using GMM without covariates for class enumeration has been questioned in the methodological literature. Obviously this practice implicitly assumes that fitting growth mixture models without covariates would recover the correct number of classes whether or not the covariates impact class membership or growth factors in the population. However, this assumption may not hold universally (Muthén, 2004). In Tofighi and Enders' (2008) study, incorporation of covariates was recognized as one of the factors that were thought to influence the extraction of the correct number of classes in the GMM context. According to Muthén (2004), auxiliary information in terms of predictors or covariates of the latent factors and latent group membership, as well as distal static outcomes of trajectory group membership (Lubke & Muthén, 2005; Petras & Masyn, 2010) can be efficiently included in a GMM analysis to obtain more accurate parameter estimates and latent class assignment. Covariates of class membership and growth factors should be included to correctly specify the model, find the proper number of classes, and correctly estimate class proportions and class membership (Muthén, 2004). Particularly, by including relevant individual-level characteristics in the model, membership in a specific trajectory group can be predicted with high probability (Nagin, 2005).

Auxiliary information may take the form of antecedents (or covariates), concurrent events, or consequences (or distal outcomes). The unconditional growth mixture model, like that specified in Equation 5 and Equation 6, can be extended in many ways based on the relation between auxiliary variables and the growth factors and/or the latent class membership. For example, covariates can enter the basic growth mixture model to explain individual differences in growth attributes. They can also be related with latent group membership.

To help understand how covariates and distal outcomes are related to a GMM analysis, one such extended growth mixture model is shown in Figure 2.1. The model consists of the following components: covariates or predictor variables (*X*), a categorical latent class variable (*C*), repeated continuous outcomes (*Y*), growth intercept (η_0) and slope (η_1), and a distal outcome variable (*Z*) as the consequences of the growth process. In terms of covariates, both time-variant and -invariant covariates (e.g., treatment and intervention effects) can be included in the GMM framework. Since the effect of time-invariant covariates is what the study will examine, time-varying covariates will not be discussed further. Time-invariant covariates can be incorporated in the GMM analysis in several ways. First, the categorical latent class variable *C* may be related to covariates X via a multinomial logistic regression model which specifies the functional relation between the probability of class membership and set of covariates X, as expressed by Equation 8 below.

$$p(C_i = k \mid \mathbf{x}_i) = \frac{\exp(\gamma_0^k + \mathbf{\Gamma}^k \mathbf{x}_i)}{\sum_{h=1}^{K} \exp(\gamma_0^h + \mathbf{\Gamma}^h \mathbf{x}_i)}, \tag{8}$$

where class *K* is the reference class and $\gamma_0^K = 0$ and $\mathbf{\Gamma}^K = 0$ for identification purposes so that the log odds of comparing class *k* to the last class *K* is

$$\log[p(C_i = k \mid \mathbf{x}_i) / p(C_i = K \mid \mathbf{x}_i)] = \gamma_0^k + \mathbf{\Gamma}^k \mathbf{x}_i. \tag{9}$$

Here $\mathbf{\Gamma}^k$ is a $1 \times q$ vector of regression coefficients denoting the effect of **x** on the log odds of membership in class *k* relative to class *K*, and γ_0^k is the logistic regression intercept for class *k* relative to class *K*. Lubke and Muthén (2007) pointed out that it was important to include in a growth mixture model the covariates that predicted class membership when examining the latent classes. The arrow from *X* to *C* in Figure 2.1 shows this type of direct relation.

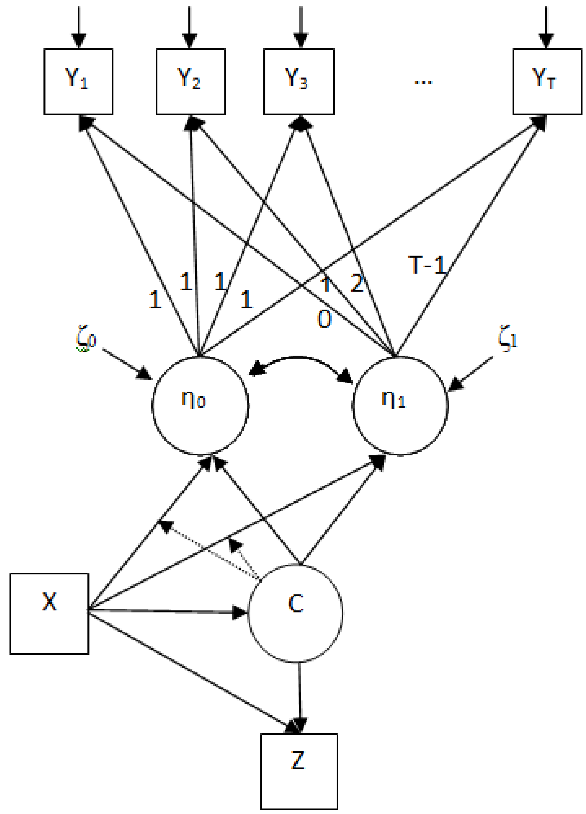


Figure 2.1. Path diagram for a general linear growth mixture model with time-invariant covariates (X) and distal outcome of change (Z)

Time-invariant covariates can also enter the growth mixture model as direct predictors of trajectory parameters. In this case, the direct effects from the covariates to the growth factors can be class-invariant or class-specific. The direct effect of class-invariant covariates on the growth factors is shown by the arrows pointing from *X* to η_0 and η_1 in Figure 2.1. These direct effects on the growth factors can be expressed in the population model of Equation 6 as

$$\boldsymbol{\eta}_i^k = \boldsymbol{\alpha}^k + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \tag{10}$$

where $\boldsymbol{\alpha}^k$ is a $q \times 1$ vector of conditional regression intercepts for $\boldsymbol{\eta}$ within class *k*, and $\boldsymbol{\Gamma}$ is a $q \times r$ matrix of regression coefficients representing the effects of $\mathbf{x}$ on $\boldsymbol{\eta}$.

When direct effects of class-specific covariates on the growth factors occur, the association of these effects and the growth factors can be expressed below

$$\boldsymbol{\eta}_i^k = \boldsymbol{\alpha}^k + \boldsymbol{\Gamma}^k \mathbf{x}_i + \boldsymbol{\zeta}_i^k , \tag{11}$$

where $\boldsymbol{\alpha}^k$ is still a $q \times 1$ vector of conditional regression intercepts for **η** within class *k*, but $\boldsymbol{\Gamma}^k$ is a $q \times r$ matrix of class-specific regression coefficients, indicating the effect of some particular explanatory variables on **η** within class *k*. The direct effects of class-specific covariates on the growth factors are indicated by the dashed arrows pointing from *X* to η_0 and η_1 in Figure 2.1.

According to Petras and Masyn (2010), when class-varying covariates are included in the model, latent classes are defined not only by heterogeneity in growth trajectories but also heterogeneity in the effect of those covariates on the growth trajectories.

In addition to covariates, it is often interesting to include distal outcomes in a GMM analysis. According to Petras and Masyn (2010), a distal outcome can be framed in one of two ways. First, a distal outcome can be seen as an additional indicator of the latent class variable (i.e., the latent class variable captures variability in the growth factors, variability in the distal outcome, and the association between the growth factors and the distal outcome). Secondly, the distal outcome can be envisioned as an outcome of latent class membership where *Z* is not included in the estimation of the GMM, and which can be used to investigate the predictive validity of the latent classes (Clark & Muthén, 2009). Examples of distal outcomes framed as a consequence of latent class membership include alcohol dependence predicted by heavy drinking trajectory classes (Muthén & Shedden, 1999) and high school dropout predicted by mathematics achievement development trajectory classes (Muthén, 2004). Given that growth is interpreted on the latent class variable, it is reasonable to allow the latent trajectory class variable to predict the distal outcome (Muthén, 2004). The effects of covariates on the distal outcome can also be included to indicate that, for each class, the probabilities of *Z* vary as a function of *X*. The arrows from *C* to Z and from *X* to *Z* in Figure 2.1 show these specific relations. The distal outcome variable can be either continuous or categorical, and the regression can be linear, logistic, or other types of generalized linear regression models depending on the form and scale of the distal outcome. For a dichotomous distal outcome scored 0 and 1, for example, the functional relation can be expressed as

$$p(z_i = 1 \mid C_i = k, \mathbf{x}_i) = \frac{1}{1 + \exp\{\tau_k - \mathbf{v}_k \mathbf{x}_i\}}, \tag{12}$$

where z_i represents a distal outcome predicted by an individual's class membership as well as his or her background characteristics (i.e., covariates), the main effect of *C* is captured by the class-varying thresholds τ_k (an intercept with its sign reversed), and $\mathbf{v}_k$ is class-varying slopes for **x**, indicating different covariate effects on *z* for different trajectory classes. The conditional probabilities of $z_i = 1$ for each class is $\frac{\exp(\tau_k)}{1+\exp(\tau_k)}$ at $\mathbf{x} = \mathbf{0}$.

2.3 Growth Mixture Modeling Estimation via the EM Algorithm

The literature is replete with a variety of estimation methods for mixture analyses. For example, maximum likelihood (ML) estimation (see, e.g., Codd & Cudeck, 2014; Harring, 2012) has been widely used to maximize the data given a particular set of parameters. Maximum likelihood estimates for the parameters can be found iteratively through using an optimization algorithm such as Newton-Raphson. As another example, the expectation-maximization (EM) optimization algorithm (Dempster, Laird, & Rubin,

1977) can be used to circumvent the heavy computation from using, for example, a method in which the integration of the loglikelihood function must be handled directly, to a method that views the estimation problem as one which can be formulated as a missing data problem. In their study, Codd and Cudeck (2014) extended the work by Harring (2012) and discussed how SAS PROC NLMIXED could be utilized to carry out ML estimation of a nonlinear random coefficient mixture model. As an alternative to ML estimation, the literature has shown an increasing rate of applications of mixture analyses using MCMC methods within a Bayesian estimation framework (see, e.g., Depaoli, 2013; Muthén & Asparouhov, 2012; Yang & Dunson, 2010). In GMM, the main difference between ML and Bayesian estimation methods is the inclusion of prior information (i.e., a prior belief about the values of model parameters) for the modeling of the growth and variance/covariance parameters (Depaoli, 2013). While different estimation methods have been developed for mixture analyses, the current paper limits the discussion to ML estimation implemented via an EM algorithm in that the method is by far most popular estimation method used in GMM analysis and it is accessible through commercial software.

The EM algorithm is an iterative procedure for finding ML estimates and is especially useful for models that can be seen as having incomplete or missing data. The EM algorithm is a broadly applicable approach since it simplifies ML estimation substantially by reformulating the given incomplete-data problem as a complete-data problem (McLachlan & Krishnan, 2008). Because class membership is considered missing, the observed data $\mathbf{y}_i$ alone in the mixture model can be treated as incomplete. Like in any other finite mixture modeling context, in GMM estimation, it is assumed that the proportion of observations falling in latent class is unknown and must be estimated along with the other parameters of the model. Therefore, the estimation of a growth mixture model consists of two parts: the estimation of parameters related to the LGM and the estimation of class proportions (Tolvanen, 2008).

The loglikelihood function corresponding to incomplete data vectors $\mathbf{y}_i$ can be written as:

$$\ln L(\boldsymbol{\varphi}, \boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(\mathbf{y}_i \mid \boldsymbol{\theta}^k), \tag{13}$$

where

$$f(\mathbf{y}_i \mid \boldsymbol{\theta}^k) = \sum_{k=1}^{K} \varphi^k f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^k), \tag{14}$$

which shows a mixture of *K* density functions where φ^k is the class proportion for class *k*. Thus, Equation 13 can be written as:

$$\ln L(\boldsymbol{\varphi}, \boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \ln f(\mathbf{y}_i \mid \boldsymbol{\theta}^k) = \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \varphi^k f^k(y_i \mid \boldsymbol{\theta}^k) \right].$$

where $\boldsymbol{\theta}^k$ is parameter estimates related to the unconditional LGM for class *k* and $\boldsymbol{\varphi}' = (\varphi^1, \ldots, \varphi^{K-1})$. The density function for class *k* is

$$f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^k) \sim N(\boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$$

where

$$\boldsymbol{\mu}^k = \boldsymbol{\Lambda}\boldsymbol{\alpha}^k$$

$$\boldsymbol{\Sigma}^k = \boldsymbol{\Lambda}\boldsymbol{\Psi}^k\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_i^k.$$

Because GMMs contain unobserved latent variable values as well as latent class membership, there is no closed-form solution for the parameter estimates (see, e.g., Kohli, 2011; Mann, 2009). Therefore, the EM algorithm can be used to obtain the GMM model parameter estimates. To identify class membership, a vector of unobservable 0/1indicators for each individual for each class, $\mathbf{c}_i = \left\{\mathbf{c}_i^1, ..., \mathbf{c}_i^K\right\}'$, can be defined as

$$c_i^k = \begin{cases} 1, & \text{if the } i\text{th subject belongs to class } k, \qquad i = 1,...,n, \quad k = 1,..., K. \\ 0, & \text{otherwise} \end{cases}$$

Thus, the loglikelihood function for complete data can be given as (see, e.g., Muthén & Shedden, 1999; Tolvanen, 2008):

$$\begin{aligned} \ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}) &= \sum_{i=1}^{n} \ln L(\boldsymbol{\theta} \mid \mathbf{y}_i, \mathbf{c}_i) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} c_i^k \{\ln(\varphi^k) + \ln[f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^k)]\}, \end{aligned} \tag{15}$$

where the inclusion of the unknown indicator variable c_i^k implies maximizing the complete-data loglikelihood. It can also be observed that in Equation 15 the loglikelihood function is comprised of two independent parts: (1) the sum of the weighted K class probabilities and (2) the sum of the weighted K density functions. Each part can be maximized separately and reconstituted in the M-step of the EM algorithm (Muthén & Shedden, 1999).

In the E-step (i.e., expectation step) of the EM algorithm, the complete-data loglikelihood function in Equation 15 is replaced by its conditional expectation function given observed data $\mathbf{y}_i$ and the current parameter estimate $\hat{\boldsymbol{\theta}}_0$ (the initial starting value for $\boldsymbol{\theta}$ on the first iteration).

$$\begin{aligned} \ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}) &\approx E\left[\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c})\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} E[c_i^k \mid \mathbf{y}, \boldsymbol{\theta}^0]\{\ln(\varphi^k) + \ln[f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^k)]\}, \end{aligned}$$

where

$$E[c_i^k \mid \mathbf{y}, \boldsymbol{\theta}^0] = \left.\frac{\varphi^k f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^0)}{\sum_{k=1}^{K} \varphi^k f(\mathbf{y}_i \mid \boldsymbol{\theta}^0)}\right|_{\boldsymbol{\theta}^0} = \varphi_i^k. \tag{16}$$

The E-step reduces to computing the posterior probabilities for each individual (i.e., the probabilities of an individual belonging to a certain class) with respect to the parameter values at the first iteration.

These posterior probabilities are then used in the M-step (i.e., the maximization step) for maximizing the conditional expectation of Equation 15. That is,

$$\begin{aligned} \ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}) &\approx E\left[\ln L(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c})\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \varphi_i^k \{\ln(\varphi^k) + \ln[f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^k)]\}, \end{aligned}$$

where the estimates of φ_i^k replace unknown indicators c_i^k. The remaining model parameters in $f^k(\mathbf{y}_i \mid \boldsymbol{\theta}^k)$ with estimates in $\boldsymbol{\theta}^k$ (e.g., $\boldsymbol{\alpha}^k$, $\boldsymbol{\Psi}^k$, and $\boldsymbol{\Theta}_i^k$) and the class probabilities $\boldsymbol{\varphi}' = (\varphi^1, \ldots, \varphi^{K-1})$ are computed. After the M-step, the algorithm returns to the E-step to calculate new posterior probabilities and then again to the M-step (Kohli, 2011). This iteration continues until the convergence criterion related to the complete-data loglikelihood is met (Harring, 2012). A known deficit of the EM algorithm is its slow rate to converge to a solution. Yet, the popularity and usefulness of the EM algorithm for GMM applications stems from its seemingly simple implementation and how reliably it can ascertain local optima through stable, uphill steps. ML via the EM algorithm is the default estimator for mixture analyses in M*plus* 7.11, which will be used to generate data and analyze replicate data sets in the upcoming simulation.

2.4 Estimation Approaches for Growth Mixture Models with Covariates

2.4.1 Conventional three-step approach

In modeling growth mixture models with covariates, the conventional method, a step-by-step approach (D'Unger, Land, & McCall, 2002; Feng, Shaw, & Silk, 2008; Fergusson & Horwood, 2002; Jo, Wang, & Lalongo, 2010; McDermott & Nagin, 2001; Nagin & Land, 1993; Nagin, Farrington, & Moffitt, 1995), is usually adopted: First, unconditional GMM analyses are conducted based on only latent class indicators to determine the number of distinct trajectory groups. Then, predicted posterior class membership probabilities are calculated and class membership is assigned to each individual based on their highest posterior class membership probabilities. Finally, relation between the assigned latent class membership and/or the growth factors and subject-specific background characteristics and/or distal outcome(s) is investigated using either the mean comparison tests or multinomial logistic regression models. This means that the model does not include covariates or distal outcomes in the unconditional growth analyses. Influences from predictor variables are taken into account subsequently in the conditional analysis. In the third step, many researchers use mean comparisons tests, such as *t*-tests, ANOVA, or chi-square tests to summarize or compare among trajectories groups. Or, they may examine the relation between growth factors and auxiliary variables using regression analysis. Researchers may also use multinomial logistic regression models to explore the relation between latent classes and auxiliary variables, such as most likely class regression (i.e., regression of most likely class membership on the covariates), probability regression (i.e., regression of an individual's logit-transformed posterior probability to be in a given class on the covariates), and probability-weighted regression (i.e., regression that is weighted by an individual's posterior probability to be in a given class) (Clark & Muthén, 2009).

2.4.2 One-step ML approach

Rather than relating covariates to the latent class variable and/or the growth factors in a separate, subsequent step as is done in the conventional approach, an alternate estimation procedure, a one-step ML approach for estimating the effects of covariates (see, e.g., Huang et al., 2010; Muthén, 2004; Nagin, 2005; Roeder et al., 1999), was recommended which include the additional variables as part of a single model estimation of developmental trajectories to allow for the simultaneous examination of the covariates' impact on the estimation of developmental trajectories and their association with the distal outcome. By including the additional variables as part of a unified model, this one-step approach solves the problem of treating most likely class membership as an exact, observed variable by taking into account the error associated with probability estimates, and allowing individuals to be fractional members of all classes (Clark & Muthén, 2009). The one-step approach considers a model for $p(\mathbf{y}_i \mid \mathbf{x}_i)$ rather than $p(\mathbf{y}_i)$. Thus, the model has the form

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} p(C_i = k \mid \mathbf{x}_i)\, p(\mathbf{y}_i \mid C_i = k)\,, \tag{17}$$

where the probability $p(C_i = k \mid \mathbf{x}_i)$ is parameterized by means of a multinomial logistic regression model expressed in Equation 8. By allowing latent class probabilities to vary with individual characteristics, it is possible to test whether and by how much a specified covariate affects probability of class membership controlling for the level of other covariates that potentially affect latent class probability estimates. The $\mathbf{\Gamma}^k$ parameters in

Equation 8 and the multinomial parameters defining $p(\mathbf{y}_i \mid C_i = k)$ will be obtained by maximizing a loglikelihood function based on $p(\mathbf{y}_i \mid \mathbf{x}_i)$ (Vermunt, 2010), which is

$$\log L = \sum_{i=1}^{n} \log p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} p(C_i = k \mid \mathbf{x}_i)\, p(\mathbf{y}_i \mid C_i = k)\,. \tag{18}$$

Distal outcome variables can also be included in the single-step approach. However, when the distal outcome has a direct effect from both a covariate and the latent class variable, the latent class model will not be affected by this direct effect (Asparouhov & Muthén, 2013).

2.4.3 Pseudo class draw approach

Pseudo class (PC) draws (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997) is one option to independently evaluate the relation between the latent class variable and the auxiliary variables without using assigned class membership (Asparouhov & Muthén, 2006; Wang et al., 2005). The first step in the PC approach is to estimate the mixture model without covariates. During this step, posterior distribution for each individual being in each of the latent classes is calculated. Then, in the second step, using this posterior distribution, multiple pseudo-class draws for each individual's class variable are generated. That is, multiple pseudo-class memberships are obtained by making multiple random draws from the discrete posterior latent class probability distribution for each individual in the sample. This second step gives each individual a chance to fall into neighboring classes (Clark & Muthén, 2009). Typically, 20 pseudo-class draws are used for each observation, which means each individual is classified 20 times (Wang et al., 2005). These multiple pseudo class draws are used as multiple imputations of each observation's class membership as if the class membership was missing. One apparent benefit of using the random draws is that they account for the uncertainty in class assignment (Asparouhov & Muthén, 2013). The next step of the PC approach is to estimate the logistic regression model with the covariates explaining latent class membership repeatedly for the multiple draws (i.e., 20 draws), and the obtained parameter estimates are averaged. That is, the subsequent analysis is performed for each random draw, and finally the class specific information associated with the auxiliary variable(s) is obtained after results are combined across draws using the multiple imputation techniques developed in Rubin (1987).

2.4.4 Three-step ML approach

To avoid all the issues mentioned above, a new three-step ML approach was proposed by Vermunt (2010). In this new approach, the unconditional growth model would first be estimated, which is exactly the same as the initial step in the conventional three-step approach. Then, a most likely class variable is defined using the highest posterior probability from the latent class posterior distribution derived from the unconditional growth mixture analysis. In the third step, the most likely class variable is used as latent class indicator variable with classification error probability taken into account. Also, in this final stage of model estimation, auxiliary variables (e.g., relevant predictors) are introduced with the measurement model (i.e., the unconditional GMM) kept fixed. It is easily seen that the big difference between the new three-step ML approach and the conventional three-step approach is in the third step where the most likely class membership variable is treated as an imperfect measurement of latent class membership analysis in the new method but not in the conventional approach. Below is a detailed description of how the new three-step ML approach works.

The most useful part of the new three-step approach in GMMs is the posterior probabilities which is a measure of an individual's likelihood of belonging to each of the k trajectory classes based on his or her longitudinal pattern of behavior $\mathbf{y}_i$ (i.e., $p(C_i = k \mid \mathbf{y}_i)$). A posterior probability can be derived via the Bayes' rule (Dias & Vermunt, 2008; Goodman, 2007; McLachlan & Peel, 2000; Vermunt, 2010) using:

$$p(C_i = k \mid \mathbf{y}_i) = \frac{\varphi^k p(\mathbf{y}_i \mid C_i = k)}{p(\mathbf{y}_i)} = \frac{\varphi^k p(\mathbf{y}_i \mid C_i = k)}{\sum \varphi^k p(\mathbf{y}_i \mid C_i = k)}. \qquad (19)$$

During the initial latent class model estimation, posterior probabilities of class membership for each subject are computed. Then, in the second step, subjects are assigned to the most likely class membership s for the most likely class variable W using the largest posterior probabilities. Classification error probability is also considered in this step. It should be mentioned here that there are two widely used classification rules, namely, the modal assignment and the proportional assignment (Vermunt, 2010). When modal assignment is considered, class assignment is hard because a subject will be classified into the class for which $p(C_i = k \mid \mathbf{y}_i)$ is largest. This is very similar to what Fraley and Raftery (2002) referred to as *hard* assignment to the class with the highest posterior probability in the context of regression mixture models. When proportional assignment is considered, subjects are treated as belonging to latent class k with probability of $p(C_i = k \mid \mathbf{y}_i)$, which is referred to as a "soft" classification. The three-step ML approach investigated in this study focuses only on the hard assignment rule. It should be noted that although one potential limitation of using the hard class assignment rule during this step is the lack of classification accuracy which might lead to biased coefficient estimates, in the next step described in detail, it can be seen that classification error probability is considered so as to obtain more accurate parameter estimates.

In the third step when covariates are added to the GMM, a relation is established between $p(W_i = s \mid \mathbf{x}_i)$ and $p(C_i = k \mid \mathbf{x}_i)$, as shown below

$$p(W_i = s \mid \mathbf{x}_i) = \sum_{k=1}^{K} p(C_i = k \mid \mathbf{x}_i)\, p(W_i = s \mid C_i = k), \tag{20}$$

which looks similar to the one-step ML approach where the model has the form of Equation 17.   Equation 20 suggests that the new method takes into account the classification error probability (i.e., $p(W_i = s \mid C_i = k)$), which makes parameter estimates more accurate. This is different from the conventional three-step approach in which classification error is not considered at the last stage of analysis. Based on the equation above, more accurate estimates of covariate effects can be obtained by treating the most likely class variable as an imperfect measurement of the latent classes. Then, the following loglikelihood function can be maximized:

$$\ln L_{ML} = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} p(C_i = k \mid \mathbf{x}_i)\, p(W_i = s \mid C_i = k), \tag{21}$$

which yields ML estimates for both $P(C_i = k \mid \mathbf{x}_i)$ and the regression coefficients (Vermunt, 2010).

Asparouhov and Muthén (2013) discussed in detail the procedures of calculating classification error probability during step two. A matrix of average class membership probabilities needs to be established first, where W_i is the most likely class variable with *s* rows and C_i is the true latent class variable with *k* columns. Within each of the most likely latent classes, the average probability of membership for the most likely latent class as well as the remaining, 'less likely' classes for the matrix are computed. A matrix of corrected average probabilities of class membership q_{sk} is subsequently derived using

$$q_{sk} = p(W_i = s \mid C_i = k) = \frac{p_{sk} \times N_s}{\sum_s^k p_{sk} \times N_s}, \tag{22}$$

where s and k stand for, respectively, the sth row (s = 1 through s) and kth column (k = 1 through k) of the matrix, and N_s represents the sample size for the most likely class on the sth row. In the third step, the most likely class variable W_i is used as latent class indicator variable with uncertainty rates prefixed at the probabilities q_{sk} (Asparouhov & Muthén, 2013). That is, the most likely latent class variable is specified as a nominal indicator of the latent class variable with logits, $\log\left(\frac{q_{sk}}{q_{Sk}}\right)$, where S is the last class. These logarithmic ratios would enter directly into the secondary statistical analysis as indicators of uncertainty (measurement error) in assigning cases to classes.

2.5 Advantages and Limitations of the Estimation Approaches

Nagin (2005) cautioned that the conventional three-step method should not be used for model estimation. Bolck et al. (2004) and Vermunt (2010) also demonstrated that the conventional three-step procedure produced biased coefficient estimates, and thus it was advocated to estimate the entire latent class regression model all at once. Clark and Muthén (2009) also discussed in detail the problems associated with some of the commonly used regression approaches mentioned above, and they pointed out that with either the mean comparison or regression methods in the third step, using the most likely class membership as an exact, observed variable was problematic. In terms of the mean comparison and the most likely class regression methods, since individuals would be assigned to the most likely class based on their highest posterior probability of being in that class, the analysis does not take into account the uncertainty of the classification. Thus, these methods are technically inappropriate for making inferences about characteristics that distinguish trajectory group membership in circumstances in which class membership is not known with certainty (Roeder et al., 1999). Similar concerns surround the probability and probability-weighted regression approaches where although probabilities of being in a class are used, errors associated with the estimated probabilities are still not taken into account, which may negatively impact the estimation of the standard errors of the regression coefficients between the posterior probabilities and auxiliary variables.

Compared with the conventional three-step method both the one-step and the more recently devised three-step ML approaches explicitly incorporate uncertainty in the derived categorical membership (McIntosh, 2013). The PC approach also takes into account classification uncertainty by using multiple random draws. In terms of the one-step ML approach, it has the advantage of taking into account the classification uncertainty by allowing individuals to be fractional members of all classes. However, one major concern may come from the impact of either the covariate variables or the distal outcome variables on the forming of latent class. That is, the latent class formed from the joint model may differ in meaning from the latent class obtained using only the indicator variables and thus may potentially change the substantive interpretation of the latent classes. Also, the method may not be practical when a large number of potential auxiliary variables are involved in the secondary analysis. Not only the prediction model but also the measurement model needs to be re-estimated when a covariate is added or deleted from the analysis, which makes exploratory work more challenging (Vermunt, 2010). Also, the decision about the number of classes in a model is hard to make considering the potential influence from including or not including covariates on class enumeration. On the other hand, simultaneously building the classification model and the prediction model may not make much sense for most applied researchers who are inclined to the idea of building the classification model first before including covariates into the analysis (Vermunt, 2010).

In terms of the new three-step ML approach, obviously, it satisfies the logic requirement of most applied researchers by following the conventional step-by-step idea. One clear advantage of this method over the conventional approach is that the most likely class membership is not treated as an exact, observed variable in the final stage analysis as was in the conventional approach. With the new approach, the most likely class variable is used with measurement error probabilities taken into account. Also, according to Asparouhov and Muthén (2013), if the class separation is good the new three-step approach has the same efficiency as the one-step approach. However, a potential problem may still exist since classification error probabilities are derived from the estimated parameters of latent class analysis without covariates, which, according to Vermunt (2010), may result in slightly underestimated standard errors.

2.6 Research on Comparing the Approaches

Studies have been conducted recently to compare the performance of various estimation approaches to incorporating covariates in mixture modeling. The main purpose of these studies was to see how efficient and reliable these methods were in terms of estimating the association between the latent class variable and auxiliary information under different conditions. For example, using simulated and real data, Clark and Muthén (2009) explored how different regression methods of relating latent class analysis results to covariates can impact estimation of auxiliary effects. Specifically, their study compared the estimates and standard errors of a regression between the most likely class membership or the posterior probabilities and a covariate using the conventional approach with those obtained from other methods: the PC method and the one-step regression approach. Results showed that the one-step approach performed the best in terms of recovering the true covariate effect. The PC method worked well when class separation was large. When class separation was not large, like the conventional regression methods, the PC method underestimated the standard errors, which is problematic because an effect may be identified as significant, when in fact, it may not be (Clark & Muthén, 2009). In another study, Vermunt (2010) compared the standard three-step procedure, the one-step approach, the BCH approach, and his proposed three-step ML approach with respect to bias in the estimates of the covariate effects and bias in the standard error estimates when covariates were included in latent class modeling. Results showed that the standard three-step approach performed poorly in the sense that its parameter estimates were severely biased downward. Both the BCH method and the three-step ML method demonstrated good parameter estimates and standard errors except when the classes were very poorly separated. It was also found that the three-step ML method was much more efficient than the BCH method in terms of the standard deviation of parameter estimates, and it was almost as efficient as the one-step estimation approach. In a very recent unpublished study by Asparouhov and Muthén (2013), the relation between a latent class variable and a predictor variable in mixture modeling was examined using different approaches under different simulation designs. Results showed that the new three-step ML approach uniformly outperformed the PC approach for analyzing the relation between a latent class variable and a covariate independently of the latent class model estimation. Also, if the class separation was substantial the three-step ML approach had the same efficiency as the one-step approach in terms of bias, mean squared error and confidence interval coverage of parameter estimates. In another recent study, Bakk, Tekle and Vermunt (2013) used both simulated and real data to investigate the association between distal outcomes and latent class variable using different methodological approaches. The results showed that the conventional three-step approach led to severely biased parameter estimates compared with other methods like the three-step ML method. However, when class separation was low, the three-step ML method underestimated the parameter estimates and their corresponding standard errors.

One limitation with these studies is that very simple latent class models for discrete responses were used. Although Asparouhov and Muthén (2013) also included more complicated models such as a growth mixture model to evaluate how well different estimation approaches performed, like most previous studies, their study included only one covariate and had a very limited number of manipulated factors and levels within those factors. Vermunt (2010) included three predictor variables in his simulation study; however, all the predictor variables were categorical. It is quite possible that in real data analytic scenarios many covariates of different types should be considered in model estimation. Bakk et al. (2013) included only distal outcome variables in their latent class analyses. Another limitation found in Asparouhov and Muthén's study with respect to GMM is that although three different types of direct effects from the auxiliary variable on the growth factors were manipulated, the impact of the covariates with various effect sizes on the new three-step estimation was not investigated.

In summary, Chapter 2 has briefly reviewed the mathematical and theoretical background of growth mixture models with auxiliary variables as well as three estimation approaches applicable for these models. To help understand the development of these complex models, the review started from the latent growth modeling procedure from which the GMM is extended by combining LGM with LCA. Similarly, the idea of including covariate(s) into LGM has been extended to GMM with auxiliary variables. The advantages of including auxiliary information to a GMM were also discussed. For example, by including relevant individual-level characteristics in the model, membership in a specific trajectory group could be predicted with high probability, which helps to correctly estimate class proportions and class membership, find the proper number of classes, and obtain more accurate parameter estimates. Also, covariates or predictors make interpretation of the growth trajectories more meaningful because of the inclusion of individual background information. Various ways of including covariate variables as well as distal outcomes into a GMM were introduced. Then, the chapter reviewed maximum likelihood estimation via the EM algorithm which is the method used in this study. Another very important section of the review is the estimation approaches for GMMs with auxiliary variables. Procedures of conventional three-step approach, one-step ML approach, and a new three-step ML approach were described in detail, whose advantages as well as limitations were also discussed. The end of the chapter reviewed research on the comparison of various estimation approaches, and limitations of previous work were noted which have lead to the idea of the current study.

In view of the limitations of previous work on examining the performance of various estimation approaches to incorporating covariates in latent class analysis, Chapter 3 aims to assess the performance of four estimation approaches (i.e., the conventional three-step approach, the one-step ML approach, the PC approach, and the three-step ML approach) for estimating covariate effects on GMMs. Specifically, covariate effect estimates on the latent class modeling will be derived using the four procedures and then compared in terms of bias estimates of the covariates effects.

Chapter 3: Methodology

This study uses Monte Carlo simulation to assess the performance of various methods used for estimating covariate effects on the latent class membership model within a growth mixture modeling framework. By using Monte Carlo simulation techniques, sample data with known population parameters are generated and the performance of the methods is evaluated under different manipulated conditions and/or model specifications. Specifically, two separate simulation studies are conducted to examine whether these methods are able to accurately estimate the relation between latent class membership and covariate(s) under two different scenarios. The experimental design of the two simulation studies is described in detail in this chapter in terms of the manipulated factors, data generation model, models used to fit the data, covariate effect definition and outcome measures used in the analyses. The software used for data simulation as well as the analysis is also discussed.

3.1 Simulation Design

This section explains the factors manipulated and why particular factor levels are considered for the two simulation studies. The manipulated factors are the same for both simulation studies and are described in detail first. Then mathematical explanation of how to manipulate degrees of class separation and covariate effect is provided. In addition, the choice of class separation levels and procedures compared in each of the two studies are discussed. Finally, the number of replications used is mentioned.

3.1.1 The same manipulated factors in the two studies

For the two simulation studies, the same fixed design characteristics include: (1) the number of time points, (2) the number of latent classes, (3) sample size, (4) proportions for the dichotomous covariate, (5) distribution of the continuous covariate, and (6) the mixing proportions of the latent classes. Both of the simulation studies include time-invariant covariates in the analyses. In addition, some simplified assumptions are made in order for the two studies to be manageable considering the complexity of the models. For example, the GMMs specified in the two studies model change for normally distributed indicator variables and assume that individual growth trajectories are linear (i.e., quadratic, higher order polynomials or nonlinear functions are not considered). It is also assumed that residual variances among indicator variables are invariant over classes (i.e., $\mathbf{\Theta}^k = \mathbf{\Theta}$ for all k) and are homoscedastic and uncorrelated (i.e., $\mathbf{\Theta} = \sigma^2 \mathbf{I}_n$), and that growth factor covariance matrices are unstructured and invariant across latent trajectory classes (i.e., $\mathbf{\Psi}^k = \mathbf{\Psi}$ for all k). Since model complexity is one factor that makes model convergence a potential issue, it has been recommended that residual variances among indicator variables as well as growth factor variances and covariances be constrained equal across classes to ensure the absence of singularities and to ensure the existence of a global solution (Hipp & Bauer, 2006; Liu, Hancock, & Harring, 2011). By adding constraints to the model, the number of free parameters to be estimated is reduced, which is expected to improve convergence in model estimation. According to Muthén (2001), mixture models are particularly sensitive to local maxima when differences in the factor variances and covariances between classes are large. Results from a simulation study by (Bauer & Curran, 2003) also showed that prediction of class membership is not more accurate when factor variances are allowed to vary than when factor variances were constrained across classes.

Tables 3.1 and 3.2 below show the fixed design characteristics and the manipulated factors for both of the simulation studies respectively. A detailed explanation of why certain level(s) are considered for use for the studies is followed immediately. It should be noted that the level(s) are selected based on a careful review of relevant simulation studies and pilot work.

Table 3.1

Fixed Factors in the Two Simulation Studies

Factor	Fixed Value
Number of repeated measures	6
Number of latent classes	2
Proportions for the dichotomous covariate	30:70
Distribution of the continuous covariate	Normal (0, 1)

Table 3.1 shows four fixed factors used for the two studies. The number of repeated measures is fixed at six assuming all individual growth trajectories in each subpopulation start and end at the same point. It is often seen in both simulation studies and substantive research of growth mixture models that the number of measurement occasions is three or more (see, e.g., Brown, 2003; Jung & Wickrama, 2008; Masyn & Brown, 2001), and it has been recommended that a minimum of three time points be used to specify a linear model (Willett, Singer, & Martin, 1998). Simon, Ercikan, and Rousseau (2012) suggested a minimum of four repeated measures to achieve more power in growth modeling. On the other hand, considering the potential issues regarding convergence or power, at least five indicators have been recommended (Muthén & Curran, 1997). Therefore, the choice of six time points seemed reasonable. As per the number of latent classes, a two-latent class model is chosen so as to keep the scope of the study manageable. For both studies,

distributions of covariates are the same. For example, proportions for the dichotomous covariate are all fixed at 30:70 and the continuous covariate has a standardized normal distribution with a mean of 0 and variance of 1.

Table 3.2

Manipulated Factors in the Two Simulation Studies

Factor	Levels
Levels of sample size	N = {500, 1,000, 5,000, and 10,000}
Mixing proportions	30:70; and 50:50
Degrees of class separation	Mahalanobis distance (MD) = {1.0, 2.0, and 3.5}
Covariate effect	Odds ratio (OR) = {1.5, 9.0}

Table 3.2 shows three manipulated factor conditions for this research. In terms of sample size, four conditions are considered which are 500, 1,000, 5,000, and 10,000. Literature review has shown various sample size ranging from 25 to 10,000. However, for latent class analysis, a sample size of 500 is considered a small sample size, especially in the low-separation condition (Vermunt, 2010). A sample size of 1,000 is selected because it is a typical sample size level used in methodological growth mixture modeling studies (see, e.g., Brown, 2003; Clark & Muthén, 2009; Kohli, 2011; Nylund et al., 2007; Tolvanen, 2008; Vermunt, 2010). The choice of a sample size of 5,000 is consistent with one of the manipulated conditions for growth mixture models by Asparouhov and Muthén (2013) whose work has been extended into this particular study, and a very large sample size of 10,000 (see, e.g., Vermunt, 2010) is added to avoid sampling fluctuation as well as to increase the convergence rate.

Results from previous studies (see, e.g., Nylund et al., 2007; Tofighi & Enders, 2008) have indicated that the mixing proportion plays an important role in growth mixture analyses. The current study will manipulate mixing proportion conditions at two levels: 30:70 and 50:50. More extreme levels such as 10:90 have led to severe convergence issues in past studies (e.g., Tolvanen, 2008) and thus will not be investigated further in this study.

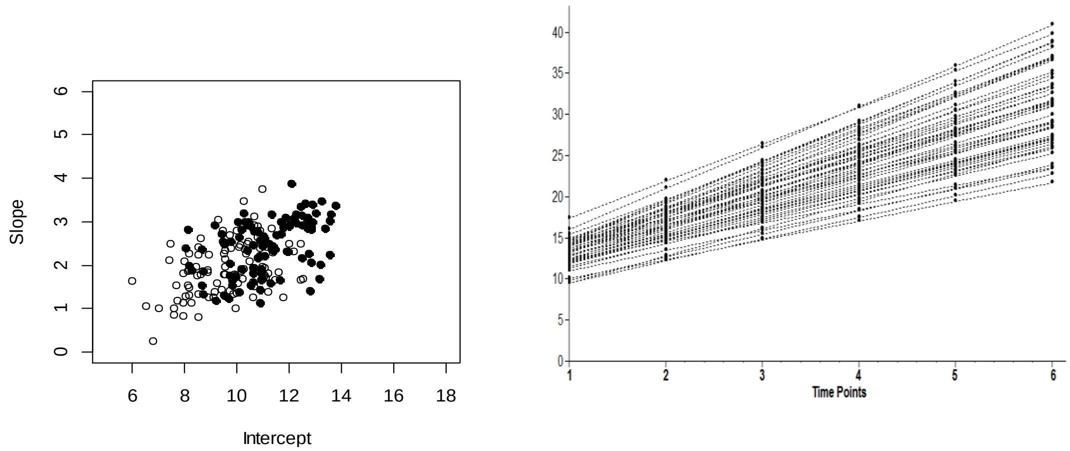
Another factor assumed to affect growth mixture analysis is class separation. As for degrees of class separation, though three levels are indicated in Table 3.2, the choice of levels for the two studies differs slightly. Specifically, all three levels are used for Simulation I whereas only two levels (i.e., MD = 2.0 and MD = 3.5) are considered for Simulation II. Mathematical explanation of the index of MD used for measuring degrees of class separation and why certain levels of class separation are selected for the two studies are provided below in Section 3.1.2.

3.1.2 Class separation and growth factor means

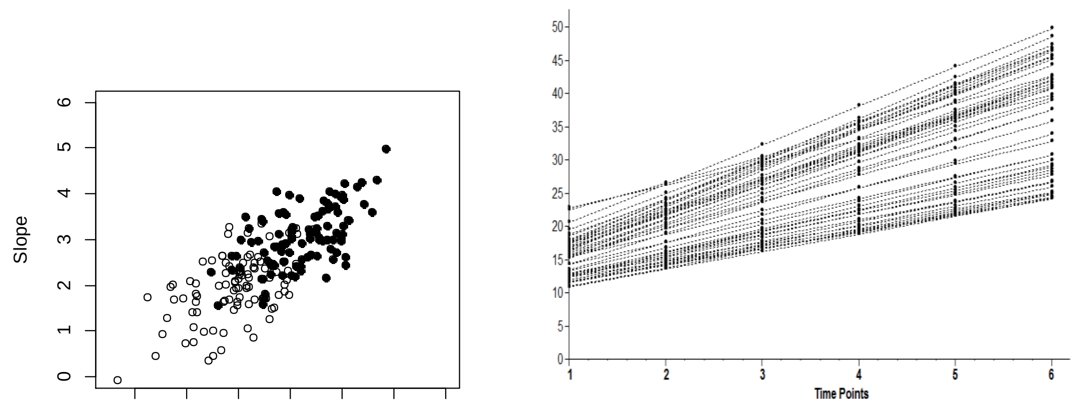
Class separation is assumed to affect the estimation approach with respect to linking covariates with the latent class variable. It has been found that the estimation accuracy of GMMs is largely affected by how well subpopulations are separated (see, e.g., Everitt, 1981; Lubke & Muthén, 2007; Tofighi & Enders, 2008). Class separation can occur at the latent level or the measured variable level (see, e.g., Tolvanen, 2008). This study will focus exclusively on class separation at the latent level for the growth parameters. Class separation in this study is measured in terms of the multivariate Mahalanobis distance (MD; Mahalanobis, 1936) between two classes and is manipulated by varying the latent growth factors (e.g., growth trajectory intercept and slope). MD between two latent classes is defined as follows:

$$MD = \sqrt{(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})' \boldsymbol{\Psi}^{-1} (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})},$$

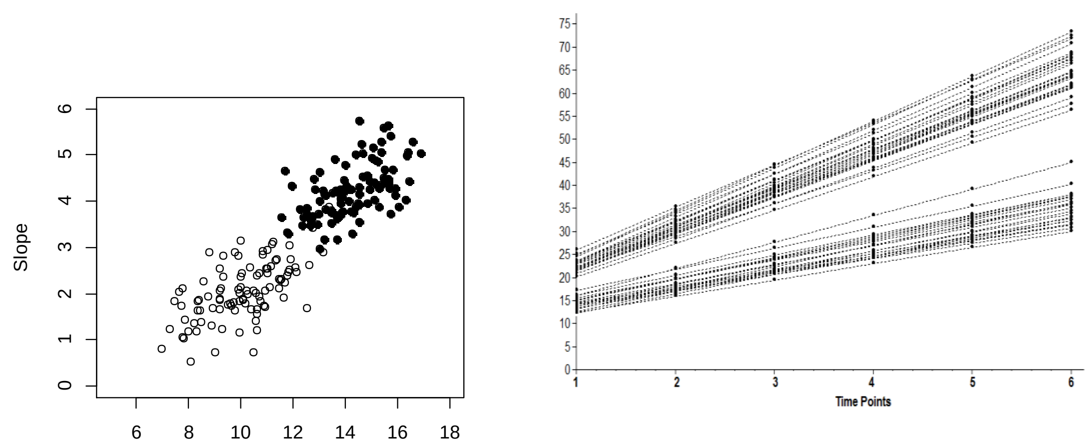
where $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ are the growth factor means for the first and second latent classes, respectively (McLachlan & Peel, 2000), and $\boldsymbol{\Psi}^{-1}$ represents the inverse of the common covariance matrix of individuals' growth parameters. In this study, the means would be the intercept and slope growth parameters for each trajectory class. Referring to previous studies and also based on exploratory analyses in a pilot study, the current research sets MDs at 1.0, 2.0, and 3.5 for Simulation I and at 2.0 and 3.5 only for Simulation II. MD values of 1.0, 2.0, and 3.5 reflect small, large, and very large trajectory separation conditions, respectively (see, e.g., Depaoli, 2013; Everitt, 1981; Lubke & Muthén, 2005; Lubke & Neale, 2006; Tolvanen, 2008; Tueller & Lubke, 2010). Small class separation (i.e., MD = 1) is not considered in Simulation II because of the extremely high non-convergence rate found in a pilot study. Figure 3.1 below shows example graphs corresponding to the three MD levels to help visually understand what degree of class separation are implied by the chosen levels of MD. It can be observed that when MD = 1, there is a great deal of overlaps between observations from the two classes, and when MD becomes a larger value such as MD = 2, the two classes are further apart from each other with some overlap, but clearly not as much as when MD = 1. There is almost no overlapping between the two classes at MD = 3.5, suggesting the two classes are well separated.



(a) MD = 1.0



(b) MD = 1.0



(c) MD = 3.5

Figure 3.1. Examples of Mahalanobis distance (MD) for two classes.

It should be added here that parameters for growth factor covariance matrices and residual variance used for both simulations are defined as:

$$\text{cov}(\boldsymbol{\eta}) = \boldsymbol{\Psi} = \begin{pmatrix} 2 & \\ .45 & .4 \end{pmatrix} \text{ and } \text{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta} = \sigma^2\mathbf{I}, \text{ where } \sigma^2 = .75.$$

It has been advocated in several simulation studies on latent growth models and GMMs that in practice, the ratio of the intercept variance to the slope variance is approximately 5:1 (see, e.g., Depaoli, 2013; Liu, 2012). In line with the consistency of this recommendation from the literature, the diagonal values in $\boldsymbol{\Psi}$ are in this ratio with the covariance set so that the correlation between the random effects is approximately 0.50.

The population values for the growth factor means (using different means of intercept and slope) under different class separation conditions are provided in Table 3.3.

Table 3.3

Growth Factor Mean Parameters under Different MDs

	MD = 1		MD = 2		MD = 3.5	
Growth Factor Mean	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Intercept	10	11.22	10	12.44	10	14.28
Slope	2	2.55	2	3.09	2	4.19

3.1.3 Manipulating covariate effect

Covariate effect size is one major manipulated factor in this research. Covariate effect size with respect to the strength of the association between the covariate(s) and class membership is manipulated using odds ratio (OR). Odds ratio estimates the change in the odds of membership in the target group (i.e., class 1) for a one unit increase in the predictor. The covariate can be either dichotomous or continuous. Two levels of OR, 1.5 and 9.0, are considered for both simulations in this study, indicating small and large effect, respectively (see, e.g., Cohen, 1988). Detailed description of these effects is to be found in Section 3.2 for Simulation I and Section 3.3 for Simulation II. In addition to covariate effect on the latent class membership, Simulation II also incorporates covariates that enter the measurement model, which, again, is discussed further in Section 3.3.

3.1.4 Procedures compared in the two simulations

The procedures compared in Simulation I are the conventional three-step approach, the one-step ML approach, the PC approach, and the new three-step ML approach. There are only three procedures compared in Simulation II which are the conventional three-step approach, the one-step ML approach, and the new three-step ML approach. The PC approach is not considered in Simulation II because of its poor performance (see discussion below) found in Simulation I.

3.1.5 Replications

For both simulations, 500 replications in each cell of the design are executed. In methodological studies focused on growth mixture modeling, the minimum number of replications has been found to be 100 (see, e.g., Asparouhov & Muthén, 2013). Many studies have used 500 replications (see, e.g., Bauer & Curran, 2003; Brown, 2003; Nylund et al., 2007), and has also been an advocated number of replications in a recent book chapter by Bandalos and Leite (2013) to ensure an accurate portrayal of the precision in the estimates.

Data are generated and analyzed using M*plus* Version 7.11 (Muthén & Muthén, 2012).

3.2 Simulation I

Simulation I examines how well the conventional three-step approach, the one-step ML approach, the PC approach, and the new three-step ML approach perform in terms of estimating covariates effects on the latent class membership independent of the measurement model where the latent classes are determined by the pattern of growth trajectories. Data are generated such that time-invariant covariates enter the growth mixture model as direct predictors of latent class membership.

3.2.1 The data generation model

In the first simulation, the form of the logistic regression function is used to model the relation between the covariates and the latent classes. Two covariates (one categorical and one continuous) are generated as predictors of an individual being in a latent class through the multinomial logistic regression equation given as

$$\pi_i^k = p(C_i = k \mid \mathbf{x}_i) = \frac{\exp(\gamma_0^k + \gamma_1^k x_{i1} + \gamma_2^k x_{i2})}{\sum_{h=1}^{K} \exp(\gamma_0^h + \gamma_1^h x_{i1} + \gamma_2^h x_{i2})},$$

where x_{i1} is a dichotomous covariate (e.g., gender) defined with values corresponding to either 0 or 1 (e.g., female = 0 and male = 1), and x_{i2} is a continuous covariate (e.g., aptitude) having a standardized normal distribution with a mean of 0 and variance of 1. The regression coefficients (i.e., γ_1^k and γ_2^k) represent the effect of covariates on the log odds of membership in class *k* relative to class *K*, and γ_0^k is the logistic regression intercept for class *k* relative to class *K*. For simplicity, interaction between the two covariates is not considered in this study. For purposes of model identification, latent class 2 will be considered the reference class, then coefficients, γ_0^2, γ_1^2, and γ_2^2 are all fixed as 0. The final logistic model can then be expressed in its logit form as:

$$logit\ (\pi_i^1) = log\ (\pi_i^1 / \pi_i^2) = \gamma_0^1 + \gamma_1^1 x_{i1} + \gamma_2^1 x_{i2}$$

A path diagram is created and is shown in Figure 3.2 to help understand the data generation model for Simulation I, where $Y_1 - Y_6$ are the six repeated measures, η_0 and η_1 are the intercept and the slope respectively, X stands for the covariates, and C is the categorical latent class variable. The arrow from X to C shows that the covariates enter the growth mixture model as predictors of latent class membership.

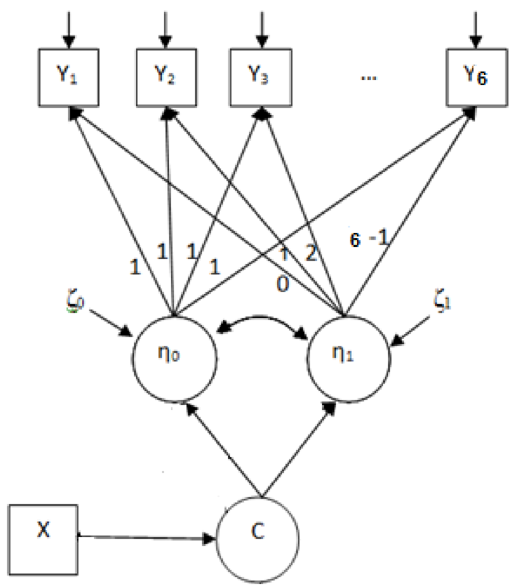


Figure 3.2. Path diagram for the data generation model for Simulation I

It should be noted that the predictors, x_1 and x_2, are generated such that the strength of the correlation between these two variables is weak to moderate positive, $\rho = 0.30$. Inducing the correlation between categorical and continuous variables in this research is to mimic the real life situation where most of the variables are correlated and independent relation between variables seldom exists. Since M*plus* software program does not include an algorithm for directly generating a categorical variable, the correlation between the dichotomous variable x_1 and the continuous variable x_2 are produced following the procedures described below.

Suppose that x_1 and x_2 follow a bivariate normal distribution with a correlation of $\rho_{x_1x_2}$ (in our case, $\rho_{x_1x_2}$ = 0.3). If x_1 is dichotomized to produce x_{1D}, then the resulting correlation between x_{1D} and x_2 can be designated as $\delta_{x_{1D}x_2} = \rho_{x_1x_2}(h/\sqrt{pq})$, where p and q are the proportions of the population above and below the point of dichotomization, respectively, and h is the ordinate of the normal probability density function at the same point (Magnusson, 1966). Values of *h* for any point of dichotomization can be found in standard tables of normal curve areas and ordinates (e.g., Cohen & Cohen, 1983, p. 521), and the sign of correlation in the equation should not change with dichotomization. Therefore, instead of using 0.3, the correlation parameter used in this study for data generation is: 0.395.

3.2.2 Covariate effect

As was mentioned earlier, covariate effect size with respect to the strength of the association between the covariates and class membership is manipulated using odds ratio (OR). Two levels of OR are set for both x_1 and x_2 as 1.5, and 9 to indicate small and large effect, respectively (see, e.g., Cohen, 1988). Therefore, four sets of covariate effects for x_1 and x_2 are manipulated, which are: 1.5 for x_1 and 1.5 for x_2, 9 for x_1 and 9 for x_2, 1.5 for x_1 and 9 for x_2, and 9 for x_1 and 1.5 for x_2. Since x_{i1} is defined with values corresponding to either 0 or 1, an odds ratio between different covariate groups and latent classes is shown as:

$$\text{Odds ratio} = \frac{p(C=1 \mid x_1=0)/p(C=2 \mid x_1=0)}{p(C=1 \mid x_1=1)/p(C=2 \mid x_1=1)} = \exp(\gamma_1),$$

where the odds of being in class 1 is approximately $\exp(\gamma_1)$ times greater for one categorical group (e.g., males) than the other (e.g., females). Levels of odds ratio are also used to manipulate the strength of the relation between x_2 and the latent class membership, although the interpretation is different from that for the dichotomous covariate. Specifically, for a one unit increase in x_2 (e.g., aptitude), it is expected to result in an approximately ($\exp(\gamma_2)$ -1) increase or decrease in the odds of being in class 1, holding x_1 constant. Regression coefficient parameters used for generating the data are provided in Table 3.4 below.

Table 3.4

Regression Coefficient Parameters for Data Generation in Simulation I

Regression Coefficient	CE = 1	CE = 2	CE =3	CE = 4
γ_1^1	0.405	2.197	0.405	2.197
γ_2^1	0.405	2.197	2.197	0.405

Note: Odds ratios are 1.5 for x_1 and x_2 at CE = 1; odds ratios are 9.0 for x_1 and x_2 at CE = 2; odds ratios are 1.5 for x_1 and 9.0 for x_2 at CE = 3; and odds ratios are 9.0 for x_1 and 1.5 for x_2 at CE = 4.

3.2.3 Summary of manipulated conditions for Simulation I

In summary, four levels of sample size, three levels of class separation, two levels of mixing proportion and four sets of covariate effects are used in the experimental design, which results in $4 \times 3 \times 2 \times 4 = 96$ cells. Since four estimation methods are examined under each of these cells, the total number of conditions is $96 \times 4 = 384$.

3.3 Simulation II

Simulation II examines how well the conventional three-step approach, the one-step ML approach, and the new three-step ML approach perform in terms of estimating covariate effects on the latent class membership when other time-invariant covariates enter the growth mixture as direct predictors of class trajectories. Therefore, the major difference between Simulation I and Simulation II in terms of data generation is that there are more covariates in Simulation II than in Simulation I and, instead of being linked only to the latent class part of the model, covariates in Simulation II are related to different parts of the growth mixture model, which has made the measurement model more complicated. As discussed earlier, time-invariant covariates can enter the growth mixture model as direct predictors of the parameters of the class trajectories, and the direct effects from the covariates to the growth factors can be class-invariant or class-specific. In the second simulation, only direct, class-specific covariates in the growth part of the model are considered while a third covariate affecting the latent class membership is also included. Thus, the total number of covariates included in Simulation II is three.

3.3.1 The data generation model

Data generation for Simulation II is more complicated than that for Simulation I. First, the form of the logistic regression function is used to model the relation between the covariate and the latent classes. Considering the model convergence issue for very complicated models, only one categorical covariate is generated as a predictor of an individual being in a latent class through the multinomial logistic regression equation given as

$$\pi_i^k = p(C_i = k \mid \mathbf{x}_i) = \frac{\exp(\gamma_0^k + \gamma_1^k x_{i1})}{\sum_{h=1}^{K} \exp(\gamma_0^h + \gamma_1^h x_{i1})},$$

where x_{i1} is a dichotomous covariate (e.g., gender). The regression coefficient, γ_1^k, represents the effect of the covariate x_{i1} on the log odds of membership in class *k* relative to class *K*, and γ_0^k is the logistic regression intercept for class *k* relative to class *K*. Again, for purposes of model identification, latent class 2 will be considered the reference class, and then coefficients, γ_0^2, and γ_1^2 are all fixed as 0. Therefore, the final logistic model expressed in its logit form is:

$$logit\ (\pi_i^1) = log\ (\pi_i^1 / \pi_i^2) = \gamma_0^1 + \gamma_1^1 x_{i1}.$$

To model the relations the covariates and growth trajectories, time-invariant covariates enter the GMM model as predictors of trajectory parameters through Equation 11. With two covariates incorporated, the associations of covariates with the growth factors can be expressed with the Level-2 model using hierarchical notation as:

$$\eta_{0i}^k = \alpha_0^k + \gamma_{01}^k x_{2i} + \gamma_{02}^k x_{3i} + \zeta_{0i}^k$$

$$\eta_{1i}^k = \alpha_1^k + \gamma_{11}^k x_{2i} + \gamma_{12}^k x_{3i} + \zeta_{1i}^k$$

where η_{0i}^k is the intercept of the true change trajectory, η_{1i}^k is the linear slope of the true change trajectory, and α_0^k and α_1^k represent population-average intercept and slope parameters within class *k*, respectively. x_2 and x_3 are Level-2 covariates, with x_2 being a dichotomous covariate defined with values corresponding to either 0 or 1 (e.g., home language: non-English = 0 and English=1), and x_3 being continuous (e.g., aptitude) which has a standardized normal distribution with a mean of 0 and variance of 1. Class-specific regression coefficients, γ_{01}^k, γ_{02}^k, γ_{11}^k and γ_{12}^k indicate the relative effect of the explanatory variables on the outcome. Specifically, γ_{01}^k and γ_{02}^k represent the effects of x_2 and x_3 on an individual's specific intercept, and γ_{11}^k and γ_{12}^k are the effects of x_2 and x_3 on an individual's specific slope parameters. Residual error terms, ζ_0^k and ζ_1^k, are bivariate normally distributed, $\zeta_{0i}^k \sim N(0, \sigma_0^2)$ and $\zeta_{1i}^k \sim N(0, \sigma_1^2)$, where σ_0^2 and σ_1^2 represent residual variances for each growth parameter, respectively, and the covariance between them is 0.45.

A path diagram is also created for the data generation model used in Simulation II (see Figure 3.3), where the dashed arrows from *X* to η_0 and η_1 indicate the class-specific covariates on the growth factors and the arrow from *X* to *C* indicates the relation between the covariates and the latent class variable.

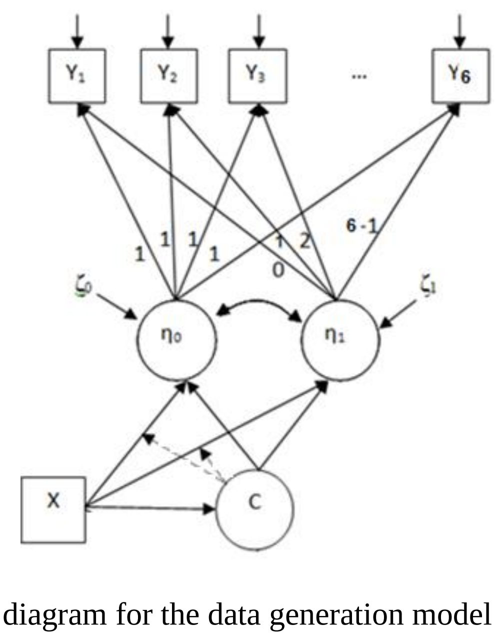


Figure 3.3. Path diagram for the data generation model for Simulation II.

As was discussed earlier, to mimic the real life situation where orthogonal relations between variables barely exist, x_1, x_2, and x_3 are generated such that the strength of the correlation between any pair of these variables is weak to moderate positive, $\rho = 0.30$. Similar to what was done in Simulation I (see Section 3.2.1), the same algorithm is used to produce the desired correlation between a dichotomous variable and a continuous variable in M*plus*.

3.3.2 Covariate effect

Like Simulation I, covariate effect size with respect to the strength of the association between the covariate and class membership is manipulated using odds ratio (OR). Two levels of OR are set for x_1 as 1.5, and 9 to indicate small and large effect, respectively (see, e.g., Cohen, 1988). Since x_{i1} is a dichotomous variable defined with values of 0 and 1, an odds ratio between different covariate groups and latent classes may be specified as:

$$\text{Odds ratio} = \frac{p(C=1 \mid x_1 = 0) / p(C=2 \mid x_1 = 0)}{p(C=1 \mid x_1 = 1) / p(C=2 \mid x_1 = 1)} = \exp(\gamma_1).$$

On the other hand, although covariate effect size with respect to the strength of the association between the covariates and growth trajectories is not examined in this study, it is manipulated in the way that Tofighi and Enders (2008) did in their study where percentage of variance explained by the covariates was used for covariate effect control. Specifically, following their example, the values of the coefficients are chosen arbitrarily such that the covariates account for 16% of the intercept and slope variation in Class 1, and 6% of the variation in the growth factors in Class 2. These proportions of explained variance roughly correspond with Cohen's (1988) effect size benchmarks in the multiple regression context (e.g., 6% and 14% approximate a medium and large effect size for R^2). The regression coefficient parameters used for generating the data for Simulation II are provided in Table 3.5, and the algorithm used for generating these regression coefficient parameters is described in Appendix A.

Table 3.5

Regression Coefficient Parameters for Data Generation in Simulation II

Latent Class	Intercept	Slope	CE = 1	CE = 2
Class 1	$\gamma_{01}^{1} = 0.336$ $\gamma_{02}^{1} = 0.500$	$\gamma_{11}^{1} = 0.232$ $\gamma_{12}^{1} = 0.200$	$\gamma_{1}^{1} = 0.405$	$\gamma_{1}^{1} = 2.197$
Class 2	$\gamma_{01}^{2} = 0.500$ $\gamma_{02}^{2} = 0.200$	$\gamma_{11}^{2} = 0.200$ $\gamma_{12}^{2} = 0.100$		

Note: Odds ratios are 1.5 for x_1 at CE = 1 and 9.0 for x_1 at CE = 2; regression coefficients of intercept and slope are chosen for Class 1 such that the covariates account for 16% of the intercept and slope variation; regression coefficients of intercept and slope are chosen for Class 2 such that the covariates account for 6% of the variation in the growth factors.

3.3.3 Two models used for Simulation II

Two models are used for Simulation II: the correctly specified model and a misspecified model. By correctly specified model, we mean that the data generation model is used for data analysis. That is, x_1 is included in the latent class and x_2 and x_3 are incorporated in the measurement part of the model. In terms of the misspecified model, only one condition is considered where the two covariates associated with the growth factors are not included in the data analysis.

3.3.4 Summary of manipulated conditions for Simulation II

In sum, four levels of sample size, two levels of class separation, two levels of mixing proportion, and two levels covariate effects on the latent class membership are included in the second simulation design, which result in 4 × 2 × 2 × 2 = 32 cells. Since three estimation methods will be examined under each of these cells, and two models are used for data analysis, the total number of conditions is 32 × 3 × 2 = 192.

3.4 Criteria for Evaluating Estimation Approaches

One might consider a method to perform well when the parameter estimates stemming from that estimation approach are unbiased and their variation is small. Therefore, the outcome measures to be compared in the current research include: (1) percent relative bias in the estimates of the covariate effects, (2) variance of the covariate effect estimates, and (3) standard error efficacy of the covariate effect estimates. In addition, estimation convergence will be examined and monitored.

One criterion to be used for evaluating the four estimation methods will be the percent relative bias for the covariate effect estimates. Bias is defined by the average difference between the population-generating covariate effect value and the parameter estimates, which is expressed as

$$\text{Bias of } \hat{\theta}_{bias} = E[\hat{\theta}] - \theta$$

where θ is the true covariate effects (population parameter) and, $\mathrm{E}[\hat{\theta}]$ is the expected covariate estimates computed from the replicate data sets within each cell of the design. A percent relative bias will be obtained by dividing the bias of a parameter estimate (i.e., estimate of a covariate effect) by the population parameter value, which is expressed as

$$\text{Relative bias of } \hat{\theta}_{rb} = \left(\frac{E[\hat{\theta}] - \theta}{\theta} \right) \times 100 .$$

Relative bias may be preferred in this situation because the magnitude of the parameter estimates in the analyses will be on different scales and thus relative bias essentially removes the scale of the parameter in its calculation putting the values on equal footing.

Variance of covariate effect estimates within each cell will also be compared. Variance is informative in that it suggests the variability of parameter estimates in the population by examining its empirical sampling distribution and is calculated as

$$\mathrm{var}(\hat{\theta}) = 500^{-1}\sum_{j=1}^{500}(\hat{\theta} - \bar{\hat{\theta}})^2 .$$

Standard error efficacy of the covariate effect estimates will be used as another criterion for estimation method comparison, which can be obtained using

$$\text{Standard Error Efficacy of } \hat{\theta} = \frac{SE(\hat{\theta})}{SD_1(\hat{\theta})},$$

where $SE(\hat{\theta})$ is the square root of the mean variance of $\hat{\theta}$ derived from the 500 replications (i.e., $SE(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^{500}(SE_j(\hat{\theta}))^2}{500}}$, where $SE_j(\hat{\theta})$ is the standard error estimates of θ for replication *j*), and $SD_1(\hat{\theta}) = SD(\hat{\theta}) \cdot \sqrt{\frac{499}{500}}$ which is the corrected sample standard deviation of 500 parameter estimates in a given cell. If the estimated standard errors computed based on an approach are accurate, the ratio of $SE(\hat{\theta})$ to $SD_1(\hat{\theta})$ should be close to 1 (Lee, 2007; Lee, Song & Poon, 2004). It should be noted that unlike the mean of the standard error estimates (i.e., $\frac{\sum_{j=1}^{500} SE_j(\hat{\theta})}{500} = \frac{\sum_{j=1}^{500}\sqrt{\mathrm{var}_j(\hat{\theta})}}{500}$), $SE(\hat{\theta})$ is an unbiased estimate of the true sampling variability. This is because the standard error estimates provided by the software programs are in fact the square root of the variances, and although taking the square root does not result in biased variance of an estimator, this nonlinear transformation causes a biased estimator of the population standard error.

Going back to the standard error efficacy of the covariate effect estimates, values greater than 1 indicate that the standard errors are overestimated, implying increase of committing Type II errors by the model whereas values less than 1 indicate that the standard errors are underestimated by the model (chance of committing Type I errors).

Using the collated data for the three evaluation criteria (i.e., relative bias, parameter estimate variance, and standard error ratio) as dependent variables, three separate repeated measures ANOVAs will be conducted to determine the statistical significance of the effects of the different levels of the manipulated factors in various covariate estimation approach conditions. In the ANOVA, all conditions are treated as the fixed effects. In summary, four main effects (i.e., sample size, degrees of class separation, mixing proportion, and covariate effects) and their interaction terms, up to three-way interaction, will be included in each of the three models used in this research.

In addition to test the statistical significance (i.e., a significant effect is claimed if *p*-value < nominal α level), in order to determine the practical significance of the effect, an effect size index, eta-square (η^2), which is defined as $\eta^2 = \frac{SS_{effect}}{SS_{total}}$ will also be assessed. An η^2 of 0.06 indicates a medium sized effect (see, e.g., Cohen, 1988) and will be used as a cutoff for practical significance with smaller values denoting impractical significance. Using the results from the factorial ANOVAs will guide which findings to focus on when reporting the results of the simulation studies.

3.5 Two Potential Issues to Address

3.5.1 Label switching

Label switching refers to the arbitrary mismatch between estimated class membership and generating class membership for simulated data in mixture modeling.

Only one of the possible permutations is the correct match and others indicate an occurrence of label switching. The occurrence of the label switching has to be detected and mismatched class membership has to be corrected before aggregating parameter estimates from multiple replications. Failing to match the correct class membership will result in the incorrect evaluation of the accuracy of the parameter estimation.

Since the current study uses ML estimation via the EM algorithm for the analysis of the growth mixture models, the label-switching issues present in Bayesian MCMC estimation (between- and within-chains) do not exist here. However, as has been pointed out in previous (see, e.g., McLachlan & Peel, 2000; Tueller, Drotar, & Lubke, 2011), the class labels are arbitrary in mixture models without previous knowledge of subpopulations. In simulation studies, parameter estimates are aggregated over replications and from replication to replication the same classes may not be labeled the same. It is critical to avoid aggregating parameter estimates over mislabeled classes. The label-switching problem can be prevented by using true parameter values starting values, making model constraints or inspecting parameter estimates after estimation. In this study, all three procedures will be implemented. In terms of model constraints, MODEL CONSTRAINT commands are included in the M*plus* syntax. For example, if we define *i1* as intercept for Class 1, *i2* as intercept for Class 2, and *s1* as slope for Class 1 and *s2* as slope for Class 2, then we can add “MODEL CONSTRAINT: i1 < i2; s1 < s2;” to the M*plus* code to make sure the labeled Class 1 does have higher intercept or slope than Class 2.

3.5.2 Convergence

In terms of the convergence, problems are regularly found in mixture model studies. Since the current study only examines the parameter recovery in converged cases, low convergence rates will undermine the evaluation of parameter recovery and subsequent factorial analysis of variance results. The distribution of estimates from limited number of replications might not represent the true sampling distribution of population parameters. Unbalanced cell sizes within the factorial design may hinder the interpretation of ANOVA results. For the two simulations, new datasets were generated and estimated until the number of replications converged for each simulation cell reaches 500. Detailed reports of the convergence rate will be presented and discussed later in Chapter 4.

It should be added that M*plus* is flexible in terms of setting starting values, number of random starts and final optimizations, and perturbation levels to mitigate problems with model convergence under the EM algorithm. In both of the simulations, true population parameters will be used as the starting values to provide efficient information for the estimation algorithm to obtain improved model convergence. In terms of number of random starts and number of final optimizations, the default for latent variable mixture analysis in M*plus* 7.11 (Muthén & Muthén, 2012) is 10 random sets of start values with two solutions with the highest log-likelihoods chosen as the starting values to be iterated until convergence is obtained or the iterative estimation is stopped due to a lack of convergence. In M*plus* syntax, the STARTS = 50 10 option will be used to change the number of starting values in the initial stage from 10 (default) to 50 and the number of final optimizations from 2 (default) to 10. In addition, perturbation level of the starting values will be changed from 5 (default) to 3. Selection of these values is made based on findings from previous studies (see, e.g., Hipp & Bauer, 2006; Li, Harring, & Macready, 2014).

Chapter 4: Results

In this chapter the results of two Monte Carlo simulations are presented. The results are based on 500 replications that achieved convergence to the global solution across all estimation algorithms under investigation. Convergence rates are reported for both of the simulations. Then, results from the impact of using various approaches for estimating covariate effects on the latent class membership under different manipulated conditions are discussed separately for each of the simulations. In order to test statistically significant effects of different methods on covariate effects estimation under the manipulated factors, several repeated measures analyses of variance (ANOVA) were conducted in SPSS (version 18.0). Specifically, percent relative bias, variance of the covariate effect estimates and standard error efficacy from using various estimation approaches were compared under the simulation conditions.

To make presentation of the results concise, a list of abbreviations of the manipulated factors and the estimation methods is used in the tables and the graphics and shown in Table 4.1.

Table 4.1

Abbreviations of the Manipulated Factors and the Estimation Methods

Factor	**Abbreviation**

Covariate Estimation Approach	A
Conventional Three-step Approach	A1
One-step ML Approach	A2
PC Approach	A3
Three-step ML Approach	A4
Model	M
Misspecified Model	M1
Correctly Specified Model	M2
Sample Size	N
Latent Class Mixing Proportion	MP
Class Separation	CS
Covariate Effect	CE

4.1 Convergence Rate

Although convergence issues were not the intended focus of this research, it is still interesting to see how well the replications converge under the various simulated conditions using different covariate effect estimation methods. Because the common problems with using the EM algorithm for fitting any type of mixture model are non-convergence or local maxima, the divergent replications in this study included replications that failed to converge to a consistent solution or converged replications with local maxima. The convergence rate for each condition was calculated using the first 500 replications. Proportion of properly converged replications for each of the estimation methods were reported for each of the two simulations at the different covariate effect sizes.

4.1.1 Convergence rate for Simulation I

Tables 4.2 – 4.5 showed the rates of converged replications for the four estimation approaches under various simulated conditions at different levels of covariate effect. One common observation from the four tables was that across all levels of covariate effect, convergence rates were 100% for all of the four estimation methods across levels of sample size and mixing proportion when class separation was at the highest level of MD = 3.5. It was also observed that the convergence rates for the PC approach and the three-step ML approach were above 95% under all the 96 simulated conditions. Eighty-nine out of 96 (92.7%) cells showed convergence rates of over 90% for the conventional three-step method, and the convergence rates for the other 7 cells ranged from 85.6% to 89.4%. An examination of the convergence rates for the one-step ML approach showed that the convergence rate was as low as 49.8% at the condition of MD = 1.0, mixing proportion of 30:70, and sample size of 500 when both x_1 and x_2 had small covariate effect (see Table 4.2), and that the convergence rate under the same combined condition was 75.0% (see Table 4.5) when covariate effect was large for x_1 and small for x_2. The low convergence rate of 50.6% from using the one-step ML approach was also observed in Table 4.2 at mixing proportion of 50:50 when class separation was at MD = 1.0, sample size was 500, and both x_1 and x_2 had small covariate effect. The low convergence rates observed in Tables 4.2 and 4.5 suggested that compared with the other three estimation approaches, the one-step ML approach was more sensitive to low class separation, small sample size and the size of covariate effect from the continuous variable (i.e., x_2). However, it was also noticed in these two tables that at MD = 1.0 the convergence rates from the one-step approach improved dramatically at both levels of mixing proportion when sample size increased.

Table 4.2

Convergence Rate with Small Covariate Effects for both x_1 and x_2 (%)

Class Separation (MD)	Mixing Proportion	Sample Size	Conventional 3-Step	1-Step ML	PC	3-Step ML
1.0	30:70	500	**88.0**	**49.8**	96.4	96.4
		1000	90.8	65.4	96.0	96.0
		5000	96.2	97.4	96.0	96.0
		10000	98.6	100	99.0	99.0
	50:50	500	92.2	**50.6**	98.4	98.4
		1000	92.6	65.8	97.2	97.2
		5000	93.2	99.0	97.6	97.8
		10000	96.4	100	97.2	97.2
2.0	30:70	500	99.6	94.0	99.2	99.2
		1000	100	99.8	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	99.2	92.4	99.2	99.2
		1000	100	99.4	100	100
		5000	100	100	100	100
		10000	100	100	100	100
3.5	30:70	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100

Note: the bolded numbers are the numbers discussed in Section 4.1.1.

Table 4.3

Convergence Rate with Large Covariate Effects for both x_1 and x_2 (%)

Class Separation (MD)	Mixing Proportion	Sample Size	Conventional 3-Step	1-Step ML	PC	3-Step ML
1.0	30:70	500	**85.6**	94.0	95.8	95.8

		1000	**88.6**	97.6	97.2	97.2
		5000	93.0	100	96.6	96.6
		10000	96.6	100	97.8	97.8
	50:50	500	90.8	94.6	96.0	96.0
		1000	92.0	99.4	94.6	94.6
		5000	91.0	100	95.8	95.4
		10000	94.0	100	97.6	97.6
2.0	30:70	500	99.4	99.4	99.6	99.6
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	99.4	100	98.0	98.0
		1000	100	100	99.2	99.2
		5000	100	100	100	100
		10000	100	100	100	100
3.5	30:70	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100

Note: the bolded numbers are the numbers discussed in Section 4.1.1.

Table 4.4

Convergence Rate with Small Covariate Effect for x_1 and Large Covariate Effect for x_2 (%)

Class Separation (MD)	Mixing Proportion	Sample Size	Conventional 3-Step	1-Step ML	PC	3-Step ML
1.0	30:70	500	**88.0**	91.2	97.2	97.2
		1000	**89.4**	98.0	97.0	97.0
		5000	95.0	100	97.4	97.6
		10000	97.6	100	98.4	98.4
	50:50	500	91.0	91.6	97.2	97.2
		1000	93.4	98.8	98.0	98.0
		5000	94.0	100	96.8	96.6
		10000	96.4	100	98.0	98.2
2.0	30:70	500	99.8	100	98.4	98.4
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	99.8	100	99.4	99.4
		1000	99.8	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
3.5	30:70	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100

Note: the bolded numbers are the numbers discussed in Section 4.1.1.

Table 4.5

Convergence Rate with Large Covariate Effect for x_1 *and Small Covariate Effect for* x_2 *(%)*

Class Separation (MD)	Mixing Proportion	Sample Size	Conventional 3-Step	1-Step ML	PC	3-Step ML
1.0	30:70	500	**86.8**	**75.0**	95.2	95.2
		1000	**86.8**	89.6	95.2	95.2
		5000	93.8	99.0	97.8	98.0
		10000	96.2	99.8	99.4	99.4
	50:50	500	91.8	83.4	96.2	96.2
		1000	91.8	94.4	97.2	97.2
		5000	90.6	99.6	95.8	95.6
		10000	94.6	100	96.0	96.2
2.0	30:70	500	99.4	99.8	99.0	99.0
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	99.2	98.8	98.6	98.6
		1000	100	100	99.6	99.6
		5000	100	100	100	100
		10000	100	100	100	100
3.5	30:70	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100
	50:50	500	100	100	100	100
		1000	100	100	100	100
		5000	100	100	100	100
		10000	100	100	100	100

Note: the bolded numbers are the numbers discussed in Section 4.1.1.

4.1.2 Convergence rate for Simulation II

As was mentioned earlier in Chapter 3, more covariates were incorporated in Simulation II than were in Simulation I and, unlike Simulation I where covariates entered only the latent class part of the model, one covariate in Simulation II was linked to latent class membership prediction while two other covariates entered the model as direct predictors of the parameters of the latent class growth trajectories. Also, in Simulation II, three instead of four estimation procedures were considered for estimation method comparison, namely, the conventional three-step procedure, the one-step ML procedure, and the new three-step ML procedure. In addition, two models were fitted in Simulation II: a misspecified model where the two covariates associated with the growth factors were not included in the analysis and the correctly specified or the true model used for data generation.

The convergence rates for the three estimation methods under the manipulated conditions are presented in Tables 4.6 and 4.7. Please note that the misspecified model was labeled as M1 and the correctly specified model was labeled as M2 in the tables. It was observed that the proportion of converged replications for the three methods was always lower for the misspecified model than for the correctly specified model for each manipulated condition. It was also observed that for all three estimation approaches the convergence rates improved for each combined condition of mixing proportion, sample size, and covariate effect when class separation was larger under both model specifications. For example, the convergence rate for the three-step ML approach was 71.2% for the correctly specified model under the condition of large covariate effect, mixing proportion of 30:70 and sample size of 500 when class separation was at MD = 2.0 (see Table 4.7). When class separation became MD = 3.5, the convergence rate increased to 99.8%. It was also observed that the convergence rates for the correctly specified model generally improved for all three methods at each combined condition of class separation, mixing proportion and covariate effect when sample size increased.

A further examination of the two tables suggested that convergence rates were affected by the size of covariate effect as well. For example, convergence rates for the misspecified model overall increased for all of the three approaches at each combined cell of class separation, mixing proportion and sample size when covariate effect increased. It was also noticed that convergence rates for the one-step ML approach also improved for the correctly specified model at each combined condition of class separation, mixing proportion and sample size when covariate effect increased. For example, the convergence rate for the correctly specified model at MD = 2.0, mixing proportion of 30:70, and sample size of 500 was 85.4% with small covariate effect (see Table 4.6), and for the same condition when covariate effect was large the convergence rate increased to 97.8% (see Table 4.7). However, the convergence rates for the three-step ML approach decreased for the correctly specified model at class separation of MD = 2.0 when covariate effect related to x_1 increased. For example, Table 4.6 showed that the convergence rate at MD = 2.0, mixing proportion of 30:70, and sample size of 500 for the correctly specified model using the three-step ML approach was 79.0% when covariate effect was small, and when covariate effect grew larger, the convergence rate decreased to 71.2% (see Table 4.7).

Table 4.6

Convergence Rate with Small Covariate Effect for x_1 *(%)*

Class Separation (MD)	Mixing Proportion	Sample Size	Conventional 3-Step		1-Step ML		3-Step ML	
			M1[1]	M2[2]	M1	M2	M1	M2
2.0	30:70	500	60.4	82.8	59.2	**85.4**	53.8	**79.0**
		1000	54.4	88.2	58.6	97.4	58.2	92.6
		5000	56.2	91.0	55.8	100	54.6	100
		10000	52.6	89.4	56.2	100	51.8	100
	50:50	500	70.1	90.8	59.2	81.8	52.8	72.4
		1000	61.4	92.0	55.8	94.2	58.0	87.8
		5000	63.6	97.0	69.2	100	52.8	100
		10000	71.3	99.8	68.8	100	52.6	100
3.5	30:70	500	61.8	90.8	88.2	95.8	89.0	95.6
		1000	71.8	91.4	96.4	99.2	96.8	99.2
		5000	87.6	91.6	100	100	100	100
		10000	96.8	97.0	100	100	100	100
	50:50	500	92.0	92.8	89.4	90.8	80.8	90.6
		1000	96.0	96.8	98.0	98.4	93.4	98.4
		5000	98.6	98.8	100	100	100	100
		10000	98.6	99.4	100	100	100	100

Note: [1] M1 is the misspecified model; [2] M2 is the correctly specified model. It is the true model used for data generation; the bolded numbers are the numbers discussed in Section 4.1.2.

Table 4.7

Convergence Rate with Large Covariate Effect for x_1 *(%)*

Class Separation (MD)	Mixing Proportion	Sample Size	Conventional 3-Step		1-Step ML		3-Step ML	
			M1[1]	M2[2]	M1	M2	M1	M2
2.0	30:70	500	67.4	83.8	96.8	**97.8**	69.8	**71.2**
		1000	79.8	85.8	98.2	98.2	62.6	75.4
		5000	79.0	98.2	100	100	61.4	75.8
		10000	75.6	100	100	100	62.6	77.4
	50:50	500	70.2	82.2	99.0	98.6	68.4	70.2
		1000	80.4	84.8	100	100	83.0	91.2
		5000	84.6	99.8	100	100	100	97.0
		10000	88.0	100	100	100	100	100
3.5	30:70	500	87.6	94.0	100	100	100	99.8
		1000	90.8	99.4	100	100	100	100
		5000	94.6	100	100	100	100	100
		10000	95.2	100	100	100	100	100
	50:50	500	87.6	97.2	100	100	100	91.8
		1000	89.6	99.8	100	100	100	100
		5000	91.8	100	100	100	100	100
		10000	95.6	100	100	100	100	100

Note: [1] M1 is the misspecified model; [2] M2 is the correctly specified model. It is the true model used for data generation; the bolded numbers are the numbers discussed in Section 4.1.2.

4.2 Results of Simulation I

In Simulation I, performance of the four covariate effect estimation procedures was investigated under 96 simulated conditions from four levels of sample size, three levels of class separation, two levels of latent class mixing proportions and four sets of covariate effects. Specifically, covariate effect parameter recovery on the latent class membership was examined and compared for the four estimation approaches in terms of relative bias, variance of covariate effect estimates and standard error efficacy of the covariate effect estimates. Results of the three indices are reported individually in three separate sections using both descriptive statistics and repeated measures ANOVA analysis. Descriptive statistics of the three outcome indices are presented by levels of covariate effect. The main effects and up to the three-way interaction effects from the repeated measures ANOVA are reported only if they were identified to be both statistically significant (*p*-value $\leq$ .05) and had an effect size of $\eta^2 \geq 0.06$. The Huynh-Feldt correction was used to adjust the degrees of freedom when the sphericity assumption was not met.

4.2.1 Results of percent relative bias in the covariate effect estimates

4.2.1.1 Descriptive statistics of percent relative bias

Percent relative bias measures how large the bias is relative to the true value of the parameter. Relative bias was used in this study because it provided a measure of the magnitude of the bias. Relative bias magnitude close to 0 indicated less biased parameter estimates. The descriptive statistics of percent relative bias for each of the estimation methods under the 96 manipulated conditions are presented in Tables 4.8 – 4.11. To facilitate interpretation, the tables are organized by levels of covariate effect. It was observed that generally for all levels of covariate effect the magnitude of percent relative bias tended to be closer to 0 for all estimation approaches under each combined condition of sample size and mixing proportion when class separation increased. For example, in Table 4.8, for the three-step ML approach, when sample size was 500 and mixing proportion was 30:70, percent relative bias values for x_1 were 223.8, 63.7, and 10.9 at MD = 1.0, MD = 2.0 and MD = 3.5 respectively. Percent relative bias values for the covariate effect estimate for x_2 under the same conditions were 21.1, 5.6 and 0.8 at MD = 1.0, MD = 2.0 and MD = 3.5 respectively. Eyeballing the four tables also suggested that for all levels of covariate effects, the PC approach and the conventional three-step approach tended to underestimate the covariate effects. Furthermore, generally for all levels of covariate effects, percent relative bias values were much closer to 0 for the one-step ML approach and the three-step ML approach than for the PC approach or the conventional approach at any combined level of condition, suggesting the former two estimation approaches produced less biased parameter estimates.

Differences in percent relative bias were observed between the estimation approaches. For example, for the one-step ML approach, at each combined condition of mixing proportion, class separation and covariate effect, percent relative bias values were closer to 0 for both x_1 and x_2 when sample size increased. Influence from the increase of sample size on percent relative bias under the same combined manipulated condition was different for the PC approach or the conventional approach from that observed with the one-step ML method. For the PC approach, the distance between the parameter estimates and the true values was getting larger especially at the lowest class separation level when sample size increased. For example, Table 4.11 showed that for the PC method, at class separation of MD = 1.0 and mixing proportion level of 30:70, the percent relative biases for the covariate effect estimate for x_1 were -83.2, -85.9, -87.1 and -87.5 corresponding to the sample size of 500, 1000, 5000, and 10000. Similarly, relative bias for the covariate effect estimate for x_2 changed from -71.8 to -83.4 when sample size increased from 500 to 10000. In a similar fashion, the conventional three-step approach showed that percent relative bias related to x_1 was further away from 0 at the first level of covariate effect when sample size increased for each combined condition of mixing proportion and class separation (see Table 4.8), although the influence from sample size was not obvious for x_2 at any combined condition. It was interesting to notice that percent relative bias values related to x_1 were extremely far from the desired value of 0 for the three-step ML method at the first and third levels of covariate effect (where either both covariates had small effect or x_1 had small effect and x_2 had large effect) when class separation was as small as MD = 1.0. For example, Table 4.8 showed that when covariate effect was small for both x_1 and x_2, magnitudes of percent relative bias for the three-step ML approach were 223.8, 313.6, 336.2, and 401.0 at sample size of 500, 1000, 5000, and 10000 respectively when mixing proportion was 30:70 and class separation was at MD = 1.0. Table 4.10 also showed that when covariate effect was small for x_1 and large for x_2, percent relative bias values for the three-step ML approach under the same conditions were 357.2, 294.4, 247.2, and 75.9 for sample size of 500, 1000, 5000, and 10000 respectively. The extreme percent relative bias values observed in these two tables suggested that the three-step ML approach was sensitive to the covariate effect size from the dichotomous variable when class separation was poor.

Table 4.8

Percent Relative Bias with Small Covariate Effects for x_1 and x_2

Conditions	Relative Bias of $\hat{\gamma}$ (%)

			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	8.7	-56.3	83.2	27.7	-54.6	-70.1	**223.8**	**21.1**
		MD = 2.0	-15.5	-36.8	33.2	-0.2	-35.0	-49.8	**63.7**	**5.6**
		MD = 3.5	3.0	-7.2	12.3	0.7	-3.2	-10.4	**10.9**	**0.8**
	50:50	MD = 1.0	-31.7	-59.6	92.0	53.6	-50.7	-69.4	178.4	61.2
		MD = 2.0	-20.5	-35.8	23.7	2.7	-35.3	-47.3	17.9	-3.7
		MD = 3.5	-2.5	-5.7	4.1	1.5	-6.5	-9.0	4.4	1.3
1000	30:70	MD = 1.0	-48.3	-63.2	70.9	53.4	-67.2	-75.7	**313.6**	90.7
		MD = 2.0	-28.1	-37.0	10.8	3.4	-47.5	-48.9	35.3	11.9
		MD = 3.5	-6.0	-6.1	1.6	1.5	-10.0	-9.4	2.4	1.4
	50:50	MD = 1.0	-48.9	-64.1	91.1	80.3	-65.9	-74.3	180.2	-7.5
		MD = 2.0	-32.3	-35.1	5.8	0.7	-45.5	-48.4	2.7	-0.3
		MD = 3.5	-6.0	-6.2	0.6	0.8	-9.5	-9.6	0.7	0.7
5000	30:70	MD = 1.0	-56.4	-63.0	59.3	42.6	-80.6	-81.0	**336.2**	71.9
		MD = 2.0	-34.9	-33.9	1.1	0.9	-51.3	-49.5	20.2	12.0
		MD = 3.5	-6.6	-6.9	0.5	0.2	-10.9	-10.5	1.5	0.6
	50:50	MD = 1.0	-61.2	-67.5	57.0	54.8	-79.0	-80.4	112.0	7.7
		MD = 2.0	-33.7	-35.2	0.2	-0.4	-47.2	-48.3	-0.5	-0.1
		MD = 3.5	-5.7	-7.2	0.8	-0.4	-9.2	-10.6	1.0	-0.4
10000	30:70	MD = 1.0	-61.9	-62.6	8.0	10.5	-83.1	-81.7	**401.0**	167
		MD = 2.0	-35.5	-34	-0.7	0.1	-52.0	-49.7	-0.5	0.2
		MD = 3.5	-7.3	-7.0	-0.2	-0.1	-11.6	-10.7	0.0	-0.1
	50:50	MD = 1.0	-62.6	-66.1	13.8	14.7	-80.6	-81.4	12.8	-9.7
		MD = 2.0	-33.4	-34.7	0.9	-0.3	-46.7	-48.2	0.4	-0.3
		MD = 3.5	-6.3	-7.0	0.2	-0.2	-9.7	-10.3	0.3	-0.2

Note: the bolded numbers are the numbers discussed in Section 4.2.1.1.

Table 4.9

Percent Relative Bias with Large Covariate Effects for x_1 *and* x_2

Conditions	Relative Bias of $\hat{\gamma}$ (%)

			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	-80.5	-82.4	15.3	15.7	-85.3	-86.9	-6.3	-63.2
		MD = 2.0	-56.0	-59.3	6.3	5.0	-66.4	-69.0	-4.2	-9.6
		MD = 3.5	-15.5	-17.5	3.0	1.9	-22.1	-24.4	2.6	1.4
	50:50	MD = 1.0	-85.7	-82.7	4.1	10.2	-91.1	-88.2	-0.1	-54.9
		MD = 2.0	-60.2	-56.0	2.3	2.6	-73.7	-68.3	7.5	-2.2
		MD = 3.5	-20.0	-17.2	1.2	1.1	-28.3	-24.1	1.9	0.9
1000	30:70	MD = 1.0	-81.2	-83.1	11.1	9.7	-86.4	-87.9	-14.4	-60.2
		MD = 2.0	-55.8	-58.3	2.4	2.2	-66.7	-69.0	-2.4	-3.3
		MD = 3.5	-16.5	-17.6	1.3	1.3	-23.2	-24.5	1.2	0.9
	50:50	MD = 1.0	-80.6	-80.3	3.4	2.7	-91.7	-89.6	56.1	-31.1
		MD = 2.0	-59.6	-56.0	1.0	0.8	-73.9	-68.9	8.6	-0.5
		MD = 3.5	-19.8	-17.4	0.5	0.4	-28.6	-24.7	1.6	0.5
5000	30:70	MD = 1.0	-80.1	-82.5	1.1	0.7	-88.4	-89.7	-10	-44.9
		MD = 2.0	-55.6	-57.9	0.1	0.1	-66.9	-69.2	-0.8	-0.9
		MD = 3.5	-17.2	-18.2	0.1	0.2	-23.9	-25.1	0.2	0.1
	50:50	MD = 1.0	-81.1	-78.9	0.9	1.1	-92.0	-89.7	89.5	-20.7
		MD = 2.0	-59.8	-56.0	0.0	0.2	-74.0	-69.0	0.0	0.1
		MD = 3.5	-20.1	-17.7	0.0	0.1	-28.9	-25.0	0.0	0.0
10000	30:70	MD = 1.0	-79.5	-81.9	-0.1	-0.1	-88.9	-90.1	-26.6	-31.8
		MD = 2.0	-55.5	-57.8	-0.1	-0.1	-66.9	-69.2	-0.5	-0.3
		MD = 3.5	-17.3	-18.2	-0.1	0.0	-23.9	-25.1	-0.1	-0.1
	50:50	MD = 1.0	-80.5	-82.4	15.3	15.7	-85.3	-86.9	-6.3	-63.2
		MD = 2.0	-56.0	-59.3	6.3	5.0	-66.4	-69.0	-4.2	-9.6
		MD = 3.5	-15.5	-17.5	3.0	1.9	-22.1	-24.4	2.6	1.4

Note: the bolded numbers are the numbers discussed in Section 4.2.1.1.

Table 4.10

Percent Relative Bias with Small Covariate Effect for x_1 *and Large Covariate Effect for* x_2

Conditions			Relative Bias of $\hat{\gamma}$ (%)							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	-32.2	-81.8	189.2	12.8	-57.8	-86.6	**357.2**	-56.8
		MD = 2.0	-20.4	-54.6	70.6	3.5	-44.8	-66.3	84.8	-5.9
		MD = 3.5	0.6	-15.3	25.4	1.3	-6.6	-22.0	26.2	0.5
	50:50	MD = 1.0	-35.1	-81.6	140.2	10.3	-57.4	-85.7	182.2	-61.0
		MD = 2.0	-26.6	-55.5	43.0	4.4	-47.2	-65.7	41.4	-9.2
		MD = 3.5	0.6	-15.0	18.5	1.1	-6.4	-21.3	17.7	0.6
1000	30:70	MD = 1.0	-51.1	-82.4	109.3	9.9	-73.5	-87.8	**294.4**	-51.3
		MD = 2.0	-36.4	-53.9	35.1	1.1	-58.6	-66.4	47.3	-2.6
		MD = 3.5	-9.3	-15.7	9.6	0.4	-17.3	-22.0	10.4	0.2
	50:50	MD = 1.0	-49.5	-81.4	70.3	8.2	-70.2	-86.6	232.8	-48.6
		MD = 2.0	-43.6	-54.2	15.8	1.3	-57.2	-65.6	14.3	-3.1
		MD = 3.5	-10.5	-15.6	4.8	1.0	-16.2	-21.1	5.0	0.6
5000	30:70	MD = 1.0	-61.2	-78.1	22.0	0.8	-87.0	-88.4	**247.2**	-24.3
		MD = 2.0	-52.4	-52.9	2.8	0.6	-69.5	-66.5	13	2.0
		MD = 3.5	-15.8	-15.3	0.0	0.2	-23.9	-22.0	1.4	0.5
	50:50	MD = 1.0	-69.4	-80.0	7.7	0.8	-84.5	-88.2	105.3	-33.6
		MD = 2.0	-51.8	-53.6	-1.1	0.3	-63.6	-65.6	-2.0	-0.7
		MD = 3.5	-14.8	-15.2	-0.8	0.3	-20.9	-21.5	-0.5	0.3
10000	30:70	MD = 1.0	-70.1	-76.5	10.9	0.5	-90.0	-88.9	**75.9**	-19.3
		MD = 2.0	-54.3	-53.0	-0.2	0.1	-70.8	-66.6	0.3	-0.2
		MD = 3.5	-16.1	-15.5	-0.1	0.0	-24.2	-22.2	0.2	0.0
	50:50	MD = 1.0	-73.6	-79.8	2.7	0.5	-87.0	-88.8	-13.9	-29.1
		MD = 2.0	-51.7	-53.5	-0.5	0.1	-63.4	-65.6	-1.1	-0.1
		MD = 3.5	-14.7	-15.3	-0.4	0.1	-20.7	-21.5	-0.4	0.1

Note: the bolded numbers are the numbers discussed in Section 4.2.1.1.

Table 4.11

Percent Relative Bias with Large Covariate Effect for x_1 *and Small Covariate Effect for* x_2

Conditions			Relative Bias of $\hat{\gamma}$ (%)							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	-88.1	0.3	-33.7	30.3	**-83.2**	**-71.8**	-1.1	-8.5
		MD = 2.0	-50.6	-36.7	-5.7	4.8	-62.8	-50.6	50.5	4.3
		MD = 3.5	-10.4	-7.7	3.5	2.1	-16.7	-11.2	7.7	1.2
	50:50	MD = 1.0	-72.4	-65.5	-7.9	44.2	-77.7	-72.1	-17.8	-23.6
		MD = 2.0	-41.7	-42.6	5.2	7.5	-52.9	-54.5	2.6	-4.8
		MD = 3.5	-7.2	-10.1	2.7	0.0	-11.4	-14.3	2.6	-0.4
1000	30:70	MD = 1.0	-74.0	-64.4	-24.0	17.8	**-85.9**	**-76.9**	3.4	-0.2
		MD = 2.0	-50.0	-36.2	-5.0	3.4	-62.9	-51.1	42.6	6.3
		MD = 3.5	-12.0	-8.1	0.9	1.2	-18.0	-11.9	3.0	0.7
	50:50	MD = 1.0	-73.3	-68.3	-4.6	14.6	-80.1	-78.1	-17.2	-21.5
		MD = 2.0	-40.5	-42.2	2.4	1.7	-53.0	-55.3	0.8	-0.8
		MD = 3.5	-8.3	-9.1	0.9	0.7	-12.5	-13.5	1.0	0.5
5000	30:70	MD = 1.0	-73.7	-65.2	-2.1	2.5	**-87.1**	**-82.9**	57.1	38.8
		MD = 2.0	-47.5	-37.9	-3.2	0.6	-62.5	-53.0	90.5	3.7
		MD = 3.5	-11.7	-8.0	0.2	0.5	-18.1	-12.3	2.7	0.8
	50:50	MD = 1.0	-72.1	-71.6	1.8	5.3	-82.6	-83.5	1.6	-17.4
		MD = 2.0	-40.2	-42.5	0.2	-0.3	-53.4	-55.9	0.0	-0.7
		MD = 3.5	-9.1	-9.7	-0.1	-0.1	-13.3	-14.2	-0.1	-0.1
10000	30:70	MD = 1.0	-71.4	-63.8	2.1	1.2	**-87.5**	**-83.4**	25.9	12.7
		MD = 2.0	-45.5	-38.0	0.0	0.3	-62.1	-53.8	0.2	0.3
		MD = 3.5	-12.0	-8.4	0.0	0.1	-18.3	-12.7	0.1	0.0
	50:50	MD = 1.0	-70.9	-71.5	2.1	2.2	-83.6	-84.5	-17.3	-22.4
		MD = 2.0	-40.1	-42.0	0.1	0.1	-53.3	-55.5	-0.2	0.1
		MD = 3.5	-9.0	-9.8	-0.1	-0.4	-13.1	-14.2	-0.1	-0.4

Note: the bolded numbers are the numbers discussed in Section 4.2.1.1.

4.2.1.2 Repeated measures ANOVA results for the percent relative bias

To better understand which factors and/or combination of factors impacted percent relative bias for the covariate effect estimates under the four estimation approaches, a repeated measures ANOVA was utilized where percent relative bias was modeled as a function of the manipulated simulation conditions. It should be mentioned that in terms of the tests of within-replications effects, the estimation approach was used as a within-replications factor because each replicated data set was exposed to each estimation approach in turn. As was mentioned before, results for up to 3-way interactions as well as the main effects were assessed and are reported in Table 4.12 only if they were identified to be both statistically significant (*p*-value $\leq$ .05) and have an effect size of $\eta^2 \geq 0.06$ (see, e.g., Cohen, 1988, p. 283; Kohli, 2010). The sphericity assumption was checked, and the Huynh-Feldt correction was used to adjust the degrees of freedom when the sphericity assumption was not satisfied.

The ANOVA results presented in Table 4.12 showed that estimation approach had a significant effect on percent relative bias of covariate effect estimates related to both x_1 and x_2. Sample size, class separation, and covariate effect had significant main between-replications effects on percent relative bias of covariate effect estimates for x_1. None of the between-replications factors showed significant effect in estimation accuracy related to covariate effect of x_2. It was observed that estimation approach had large effect sizes of $\hat{\eta}^2 = 0.46$ for x_1 and $\hat{\eta}^2 = 0.64$ for x_2, indicating that estimation approach had a large impact on the accuracy of covariate effect parameter estimates. An effect size of $\hat{\eta}^2 = 0.27$ (related to x_1) for the main effect of covariate effect suggested that estimation accuracy for the dichotomous covariate effect was greatly influenced by the levels of covariate effect manipulated. Significant two-way interaction effects for x_1 were identified for A × CS ($\hat{\eta}^2 = 0.25$), N × CE ($\hat{\eta}^2 = 0.10$), and CS × CE ($\hat{\eta}^2 = 0.28$). Interaction effects from A × CS was also found significant for relative bias related to x_2 (

$\hat{\eta}^2 = 0.20$). A × CS × CE was the only significant three-way interaction effect with an effect size of $\hat{\eta}^2 = 0.10$ and it was related to x_1. No significant three-way interaction effect was found for x_2.

Table 4.12

ANOVA Results of Manipulated Factors on the Percent Relative Bias

Source	x_1			x_2		
	F Value	*p*-value	η^2	F Value	*p*-value	η^2
Within-Replications Effects[1]						
A	701.257	<.001	0.46	911.395	<.001	0.64
A × CS	188.106	<.001	0.25	141.512	<.001	0.20
A × CS × CE	24.828	<.001	0.10			
Between-Replications Effects						
N	28.612	<.001	0.06			
CS	43.073	<.001	0.06			
CE	123.824	<.001	0.27			
N × CE	15.572	.003	0.10			
CS × CE	63.233	<.001	0.28			

Note: [1] the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary. A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion.

For the main effects, Tukey's HSD procedure was used for comparing pairs of means for the main effects of sample size, class separation and covariate effects for x_1, and the means for groups in homogeneous subsets were displayed in Tables 4.13 – 4.16. Table 4.13 showed that when sample size increased, percent relative bias of the covariate effect estimates for x_1 tended to depart from 0 (from 1.7 to -18.7). Significant changes in relative bias were found when sample size increased from 500 to 1000 and from 5000 to

10000. Table 4.14 or Table 4.15 showed no consistent pattern of change in percent relative bias across levels of either class separation or covariate effect.

Table 4.13

Pairwise Comparisons among Levels of N for Percent Relative Bias for x_1

N	Sample Size	Subset		
		1	2	3
1	24			1.7
2	24		-6.1	
3	24		-12.0	
4	24	-18.7		

Table 4.14

Pairwise Comparisons among Levels of CS for Percent Relative Bias for x_1

CS	Sample Size	Subset	
		1	2
1	32		-1.8
2	32	-19.3	
3	32		-5.3

Table 4.15

Pairwise Comparisons among Levels of CE for Percent Relative Bias for x_1

CE	Sample Size	Subset	
		1	2
1	24		9.0
2	24	-26.6	
3	24		4.5
4	24	-22.0	

Graphics were made to help to visually and directly examine the identified significant interaction effects. As was found in the ANOVA analysis, the two-way within by between interaction effects from A × CS had a significant effect on percent relative bias related to both x_1 and x_2. A comparison of Figure 4.1 and Figure 4.2 showed clearly that for both x_1 and x_2, when class separation was larger, percent relative bias from using all estimation methods was closer to the desired value of 0, which was consistent with what was observed earlier in the descriptive statistics. Percent relative bias values were closest to 0 at MD = 3.5 and furthest from 0 at MD = 1.0 for all estimation approaches. It was also observed that compared with the conventional three-step approach (A1) and the PC approach (A3), both the one-step ML approach (A2) and the three-step ML approach (A4) had percent relative bias values around 0 at MD = 2.0 and MD = 3.5. It was observed in Figure 4.1 that percent relative bias for the three-step ML approach was far away from 0 at MD = 1.0, suggesting that effect estimation for the dichotomous covariate using the three-step ML approach was biased when class separation was poor.

Graphics created for the significant two-way interactions for CE × N (Figure 4.3) and CE × CS (Figure 4.4) showed how between-replications interaction effects affected covariate effect estimate accuracy related to x_1. Figure 4.3 showed no consistency in the change of percent relative bias across the levels of covariate effect when sample size increased, although it did show that at sample size 500 percent relative bias was closer to 0 with the increase of covariate effect levels. However, it should be noted that the change from a lower to a higher covariate effect level did not necessarily mean the change of the size of covariate effect. It was just a change of conditions. In this case, it simply meant that when sample size was 500, relative bias related to x_1 had the largest distance away from 0 when covariate effect was small for both x_1 and x_2, and relative bias related to x_1 was closest in distance from 0 when covariate effect was large for x_1 and small for x_2. When covariate effect was large for both x_1 and x_2, relative bias magnitudes related to

x_1 were very close between sample sizes of 500 and 5000, and between sample sizes of 1000 and 10000. At sample size of 1000 and 10000, barely any change in percent relative bias was observed across the levels of covariate effect. Figure 4.4 showed the two-way interaction of CE × CS on the percent relative bias related to x_1. It was observed that for all levels of covariate effect, the percent relative biases were the largest in terms of their absolute magnitude at MD = 1.0, and closest to 0 at MD = 3.5, suggesting that large class separation resulted in less biased parameter estimates for any level of covariate effect. It was also observed that at MD = 2.0 and MD = 3.5, covariate effects for x_1 were underestimated for all levels of covariate effect. At MD = 1.0, covariate effects were overestimated for x_1 at the first and the third level of covariate effects and were underestimated at the second and the fourth level of covariate effects.

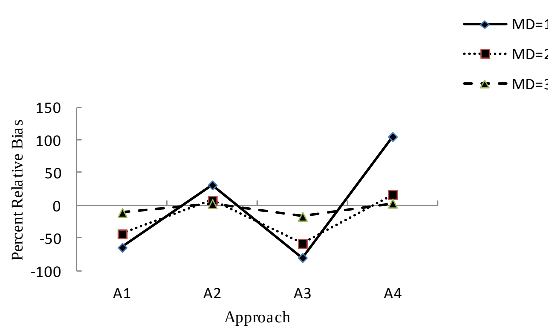


Figure 4.1. A × CS on percent relative bias related to x_1

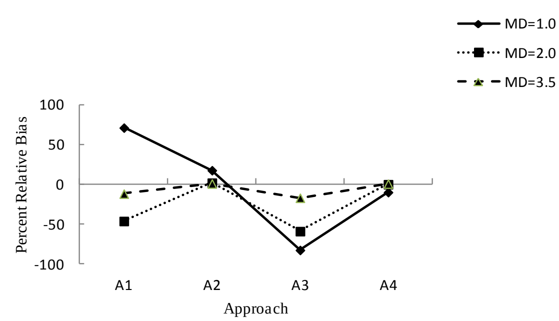


Figure 4.2. A × CS on percent relative bias related to x_2

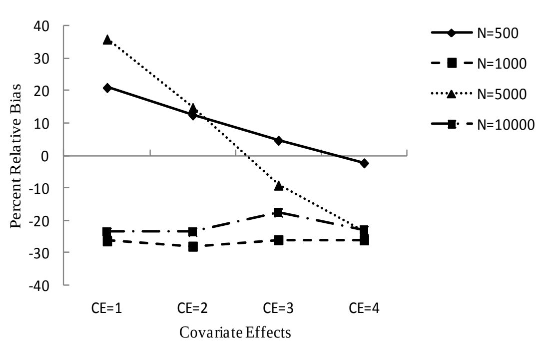


Figure 4.3. CE × N on percent relative bias

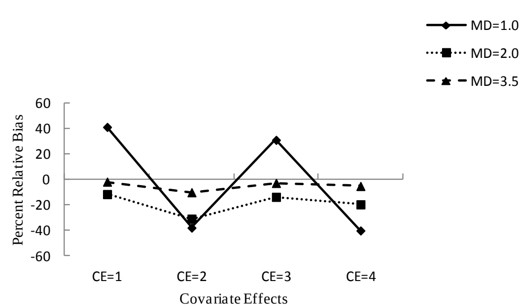


Figure 4.4. CE × CS on percent relative bias

related to x_1 related to x_1

For the three-way interaction effect of A × CE × CS on percent relative bias of effect covariate estimates related to x_1, four two-way interaction effects of A × CS were graphed for each level of covariate effect (see Figures *4.5 – 4.8*). The figures showed that for all levels of covariate effect when class separation grew larger, percent relative bias from using all estimation methods was closer to the desired value of 0, and when class separation was very large at its highest level of MD = 3.5, all estimation approaches were at their best performance in terms of covariate effect estimate accuracy for x_1. It was also observed that compared with the conventional three-step approach and the PC approach, when covariate effect was small for both x_1 and x_2, or when covariate was small for x_1 and large for x_2, the three-step ML approach lead to extreme percent relative bias values far away from 0 at MD = 1.0, indicating that parameter estimation related to the dichotomous covariate was severely affected for the three-step approach when class separation was poor and covariate effect from the dichotomous variable was small. Covariate effect estimates related to x_1 were more accurate for the one-step approach and the three-step ML approach than for the other two approaches at any class separation level when covariate effect was large for both x_1 and x_2, or when covariate effect was large for x_1 and small for x_2.

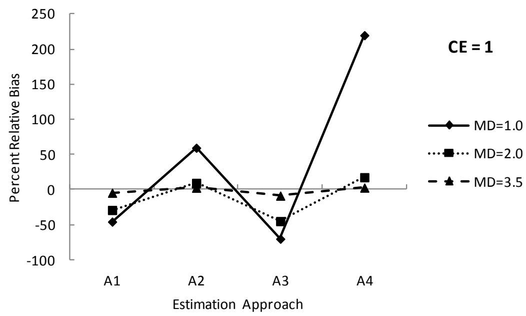


Figure 4.5. A × CS on percent relative bias for x_1 at CE=1

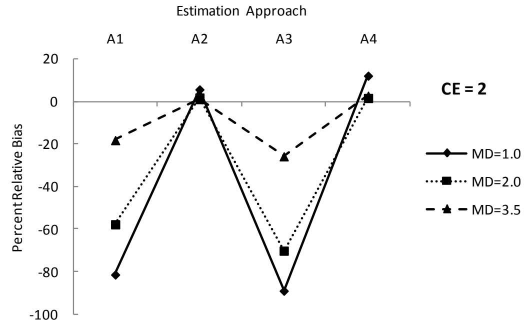


Figure 4.6. A × CS on percent relative bias for x_1 at CE=2

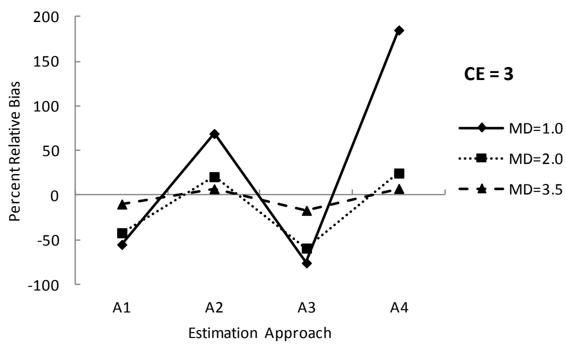


Figure 4.7. A × CS on percent relative bias for x_1 at CE=3

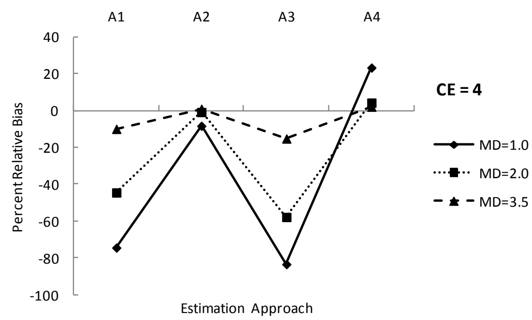


Figure 4.8. A × CS on percent relative bias for x_1 at CE=4

4.2.2 Results of variance of covariate effect estimates

4.2.2.1 Descriptive statistics of variance of covariate effect estimates

Compared to bias which indicates how close on average the estimates were to the true value, variance of covariate effect estimate suggests how much the parameter estimates change across the sample replications. It is assumed that the decrease of one of them is at the expense of increase of the other because variance of parameter estimates uses the mean of the estimates for each cell instead of the true value in measuring parameter estimate variability.

Tables 4.16 – 4.19 followed showed variances of covariate effect estimates associated with x_1 and x_2 at the four levels of covariate effect. As expected, the results indicated that variances of covariate effect estimates for the conventional three-step approach and the PC approach were generally smaller than those for the one-step ML approach or the three-step ML approach at each combined level of sample size, mixing proportion, class separation and covariate effect. When covariate effect was small for both x_1 and x_2 (see Table 4.16), variance of the covariate effect estimates related to x_1 ranged from 0.002 to 0.085 while for x_2 ranged from 0.000 to 0.019 when the conventional three-step procedure was used. With the PC method, variance of the covariate effect estimates ranged from 0.001 to 0.047 for x_1 and from 0.000 to 0.010 for x_2. It was also observed in Table 4.16 that the new three-step ML approach had the largest range of variance from 0.002 to 32.896 for the covariate effect estimates for x_1 as well as the largest range in variance from 0.001 to 58.227 for the covariate effect estimates for x_2. Variances of parameter estimates obtained using the one-step approach when x_1 and x_2 both had small covariate effect ranged from 0.002 to 1.983 for x_1 while for x_2 they ranged from 0.001 to 1.778.

A similar pattern was observed in Tables 4.17 – 4.19 for the other three levels of covariate effects related to x_1 and x_2, which suggested that compared with the conventional three-step approach and the PC approach, the one-step ML approach and the three-step ML approach resulted in more variability in terms of covariate effect estimation. Descriptive statistics in the four tables also showed that the largest variance values were found with MD = 1.0 for any estimation approach used, meaning that when class separation was poor, covariate effect estimation had more variability no matter which estimation approach was used.

Table 4.16

Variance with Small Covariate Effects for x_1 and x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	**0.085**	**0.019**	1.472	0.842	0.012	0.004	10.993	5.098
		MD = 2.0	0.039	0.012	0.088	0.026	0.020	0.006	0.2543	0.042
		MD = 3.5	0.049	0.012	0.060	0.013	**0.047**	**0.010**	0.059	0.013
	50:50	MD = 1.0	0.041	0.010	0.907	1.010	0.014	0.004	11.527	49.252
		MD = 2.0	0.032	0.012	0.081	0.036	0.018	0.006	0.071	0.029
		MD = 3.5	0.037	0.011	0.040	0.013	0.033	0.010	0.043	0.013
1000	30:70	MD = 1.0	0.024	0.008	0.644	1.265	0.006	0.002	**32.896**	**58.277**
		MD = 2.0	0.022	0.007	0.049	0.013	0.010	0.003	0.137	0.022
		MD = 3.5	0.027	0.006	0.031	0.006	0.024	0.005	0.033	0.007
	50:50	MD = 1.0	0.020	0.007	1.414	**1.778**	0.006	0.002	9.245	0.190
		MD = 2.0	0.018	0.005	0.039	0.011	0.010	0.003	0.043	0.013
		MD = 3.5	0.019	**0.005**	0.022	0.006	0.018	0.005	0.022	0.006
5000	30:70	MD = 1.0	0.029	0.007	**1.983**	1.273	0.002	0.001	10.336	0.697
		MD = 2.0	0.006	0.001	0.012	0.002	0.003	0.001	0.024	0.004
		MD = 3.5	0.005	0.001	0.006	0.001	0.005	0.001	0.007	0.001
	50:50	MD = 1.0	0.019	0.003	1.654	1.246	0.001	0.001	4.518	0.185
		MD = 2.0	0.004	0.001	0.005	0.001	0.001	0.000	0.005	0.001
		MD = 3.5	0.004	0.001	0.004	0.001	0.004	0.001	0.005	0.001
10000	30:70	MD = 1.0	0.019	0.003	0.087	0.103	0.001	0.000	11.374	22.042
		MD = 2.0	0.003	0.001	0.006	0.001	0.001	0.000	0.008	0.001
		MD = 3.5	0.003	0.001	0.003	0.001	0.003	0.000	0.004	0.001
	50:50	MD = 1.0	0.008	0.002	0.058	0.061	0.001	0.000	1.456	0.070
		MD = 2.0	0.002	0.000	0.004	0.001	**0.001**	0.000	0.005	0.001
		MD = 3.5	**0.002**	**0.000**	**0.002**	**0.001**	0.002	**0.000**	**0.002**	**0.001**

Note: the bolded numbers are the numbers discussed in Section 4.2.2.1.

Table 4.17

Variance with Large Covariate Effects for x_1 and x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	0.077	0.032	3.721	3.104	0.017	0.011	21.128	0.477
		MD = 2.0	0.128	0.032	0.539	0.233	0.041	0.013	3.543	0.281
		MD = 3.5	0.171	0.043	0.247	0.066	0.119	0.030	0.997	0.082
	50:50	MD = 1.0	0.063	0.027	3.320	2.948	0.023	0.009	17.813	0.227
		MD = 2.0	0.060	0.027	0.364	0.234	0.028	0.015	1.287	0.359
		MD = 3.5	0.089	0.042	0.119	0.061	0.067	0.029	0.141	0.075
1000	30:70	MD = 1.0	0.040	0.021	2.046	1.600	0.008	0.006	15.816	0.283
		MD = 2.0	0.050	0.012	0.246	0.088	0.017	0.006	1.075	0.130
		MD = 3.5	0.072	0.021	0.102	0.029	0.048	0.015	0.120	0.036
	50:50	MD = 1.0	0.042	0.017	2.127	1.482	0.014	0.005	13.168	0.208
		MD = 2.0	0.029	0.011	0.154	0.092	0.014	0.006	0.232	0.161
		MD = 3.5	0.043	0.020	0.058	0.030	0.033	0.014	0.069	0.036
5000	30:70	MD = 1.0	0.101	0.019	0.301	0.145	0.004	0.003	19.276	0.467
		MD = 2.0	0.013	0.002	0.052	0.019	0.004	0.001	0.110	0.029
		MD = 3.5	0.018	0.005	0.025	0.007	0.012	0.003	0.029	0.008
	50:50	MD = 1.0	0.021	0.009	0.145	0.120	0.004	0.003	6.918	0.238
		MD = 2.0	0.005	0.002	0.028	0.014	0.003	0.001	0.044	0.027
		MD = 3.5	0.008	0.004	0.012	0.005	0.006	0.003	0.013	0.006
10000	30:70	MD = 1.0	0.049	0.013	0.120	0.062	0.002	0.002	23.442	0.208
		MD = 2.0	0.006	0.001	0.027	0.009	0.002	0.001	0.042	0.014
		MD = 3.5	0.009	0.002	0.013	0.003	0.006	0.001	0.014	0.004
	50:50	MD = 1.0	0.018	0.007	0.066	0.053	0.003	0.002	0.911	0.209
		MD = 2.0	0.003	0.001	0.013	0.007	0.001	0.001	0.021	0.015
		MD = 3.5	0.004	0.002	0.006	0.003	0.003	0.001	0.007	0.003

Table 4.18

Variance with Small Covariate Effect for x_1 and Large Covariate Effect for x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	0.057	0.038	1.665	3.781	0.013	0.010	17.019	0.548
		MD = 2.0	0.039	0.025	0.176	0.219	0.016	0.012	0.549	0.291
		MD = 3.5	0.067	0.041	0.088	0.057	0.050	0.026	0.097	0.075
	50:50	MD = 1.0	0.032	0.032	1.108	3.456	0.009	0.010	9.665	0.364
		MD = 2.0	0.029	0.028	0.116	0.256	0.014	0.015	0.107	0.323
		MD = 3.5	0.048	0.038	0.059	0.052	0.039	0.028	0.063	0.063
1000	30:70	MD = 1.0	0.029	0.024	0.964	1.864	0.005	0.006	14.159	0.963
		MD = 2.0	0.023	0.013	0.077	0.082	0.009	0.006	0.134	0.119
		MD = 3.5	0.035	0.019	0.046	0.025	0.027	0.014	0.049	0.032
	50:50	MD = 1.0	0.026	0.020	0.302	1.422	0.005	0.006	15.263	16.758
		MD = 2.0	0.015	0.011	0.052	0.087	0.008	0.006	0.059	0.136
		MD = 3.5	0.022	0.017	0.033	0.027	0.022	0.014	0.036	0.031
5000	30:70	MD = 1.0	0.036	0.018	0.063	0.134	0.001	0.003	8.458	0.281
		MD = 2.0	0.007	0.002	0.019	0.017	0.002	0.001	0.031	0.027
		MD = 3.5	0.008	0.004	0.010	0.006	0.006	0.003	0.011	0.007
	50:50	MD = 1.0	0.015	0.010	0.041	0.126	0.002	0.004	6.493	9.813
		MD = 2.0	0.005	0.002	0.014	0.013	0.002	0.001	0.018	0.022
		MD = 3.5	0.006	0.003	0.008	0.005	0.005	0.003	0.008	0.006
10000	30:70	MD = 1.0	0.012	0.013	0.028	0.060	0.000	0.001	2.966	0.201
		MD = 2.0	0.004	0.001	0.011	0.008	0.001	0.001	0.015	0.013
		MD = 3.5	0.004	0.002	0.006	0.003	0.004	0.001	0.006	0.003
	50:50	MD = 1.0	0.007	0.007	0.020	0.055	0.001	0.003	0.094	0.191
		MD = 2.0	0.003	0.001	0.007	0.007	0.001	0.001	0.010	0.012
		MD = 3.5	0.003	0.002	0.004	0.002	0.003	0.001	0.004	0.003

Table 4.19

Variance with Large Covariate Effect for x_1 and Small Covariate Effect for x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	0.032	0.033	1.416	0.730	0.026	0.004	17.348	0.255
		MD = 2.0	0.119	0.012	0.413	0.029	0.042	0.006	16.648	0.040
		MD = 3.5	0.131	0.013	0.202	0.016	0.094	0.012	0.942	0.016
	50:50	MD = 1.0	1.257	0.021	1.945	0.506	0.027	0.003	8.891	0.092
		MD = 2.0	0.064	0.010	0.487	0.108	0.035	0.006	2.644	0.033
		MD = 3.5	0.071	0.012	0.087	0.015	0.058	0.011	0.101	0.015
1000	30:70	MD = 1.0	2.344	0.030	1.228	0.280	0.013	0.002	16.402	0.306
		MD = 2.0	0.041	0.006	0.288	0.013	0.016	0.003	8.313	0.020
		MD = 3.5	0.055	0.007	0.084	0.008	0.041	0.006	0.111	0.008
	50:50	MD = 1.0	0.050	0.006	1.313	0.250	0.017	0.002	6.561	0.183
		MD = 2.0	0.027	0.005	0.114	0.014	0.016	0.003	0.223	0.016
		MD = 3.5	0.035	0.006	0.043	0.007	0.029	0.005	0.048	0.007
5000	30:70	MD = 1.0	0.115	0.009	0.585	0.018	0.006	0.000	19.563	0.465
		MD = 2.0	0.008	0.001	0.043	0.003	0.004	0.001	11.542	0.004
		MD = 3.5	0.013	0.001	0.018	0.001	0.010	0.001	0.024	0.002
	50:50	MD = 1.0	0.032	0.003	0.503	0.053	0.008	0.000	6.968	0.292
		MD = 2.0	0.005	0.001	0.018	0.002	0.003	0.000	0.033	0.003
		MD = 3.5	0.006	0.001	0.007	0.001	0.005	0.001	0.008	0.002
10000	30:70	MD = 1.0	0.047	0.005	0.506	0.005	0.003	0.000	11.888	0.179
		MD = 2.0	0.006	0.001	0.030	0.001	0.002	0.000	0.052	0.002
		MD = 3.5	0.007	0.001	0.010	0.001	0.005	0.001	0.011	0.001
	50:50	MD = 1.0	0.017	0.003	0.219	0.010	0.006	0.001	0.674	0.012
		MD = 2.0	0.003	0.000	0.010	0.001	0.002	0.000	0.015	0.002
		MD = 3.5	0.003	0.001	0.004	0.001	0.002	0.000	0.004	0.001

4.2.2.2 Repeated measures ANOVA results for the variance of covariate effects estimates

As was done with percent relative bias, repeated measures ANOVA was used to identify factors and/or combination of factors that had significant impact on the variance of covariate effect estimates under the four estimation approaches. Variance of covariate effect estimates was modeled also as a function of the manipulated simulation conditions. Estimation approach was used as a within-replications variable and results for up to 3-

way interactions as well as the main effects were assessed and reported in Table 4.20 if they were identified to be both statistically significant (p-value $\leq$ .05) and have an effect size of $\eta^2 \geq 0.06$. The sphericity assumption was checked, and the Huynh-Feldt correction was considered to adjust the degrees of freedom if necessary.

The ANOVA results presented in Table 4.20 showed that estimation approach, sample size, mixing proportion and class separation had significant main effects on variance of covariate effect estimates related to x_1. Estimation approach and class separation had very large effect sizes of $\hat{\eta}^2 = 0.28$ and $\hat{\eta}^2 = 0.57$ respectively on the effect estimate variability. Only class separation was identified as significant main effect on variance of parameter estimates related to x_2 ($\hat{\eta}^2 = 0.12$). Significant two-way interaction effects for x_1 were found for A × CS ($\hat{\eta}^2 = 0.38$), N × CS ($\hat{\eta}^2 = 0.06$), and MP × CS ($\hat{\eta}^2 = 0.06$), and A × CS was also found significant for x_2 ($\hat{\eta}^2 = 0.08$). Only one three-way interaction effect (A × CS × CE) related to x_2 was found significant with an effect size of $\hat{\eta}^2 = 0.15$.

Table 4.20

ANOVA Results of Manipulated Factors on the Variance of Covariate Effects Estimates

Source	x_1			x_2		
	F Value	*p*-value	η^2	F Value	*p*-	η^2
Within-Replications						
A	157.593	<.000	0.28			
A × CS	108.551	<.000	0.38	3.137	.010	0.08
A × CS × CE				1.981	.028	0.15
Between-Replications						
CS	154.636	<.000	0.57	4.748	.022	0.12
N	10.619	<.000	0.06			
MP	33.534	<.000	0.06			
N × CS	5.247	.003	0.06			
MP × CS	15.358	<.000	0.06			

Note: [1] the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary. A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion.

Pairs of means for the main effects of sample size and class separation were compared for the variances of covariate effect estimates using Tukey's HSD procedure which was not used for the main effect of mixing proportion with only two levels. Means for groups in homogeneous subsets were displayed for the main effects of sample size and class separation in Tables 4.21 – 4.23. It was observed in Table 4.21 that as sample size increased, variance of effect estimate for x_1 grew smaller from 1.733 to 0.569. Similarly, as class separation was larger, variance of effect estimates for x_1 was smaller from 3.220 to 0.050 (Table 4.22). Table 4.23 showed that for x_2, when class separation was large at MD = 3.5, variance of parameter estimate was at the lowest value of 1.042.

Table 4.21

Pairwise Comparisons among Levels of N for Variance of Parameter Estimates Related to x_1

N	Sample Size	Subset		
		1	2	3
1	24			1.733
2	24		1.554	1.554
3	24	1.048	1.048	
4	24	0.569		

Table 4.22

Pairwise Comparisons among Levels of CS for Variance of Parameter Estimates Related to x_1.

CS	Sample Size	Subset	
		1	2
1	32		3.220
2	32	0.406	
3	32	0.050	

Table 4.23

Pairwise Comparisons among Levels of CS for Variance of Parameter Estimates Related to x_1

CS	Sample Size	Subset	
		1	2
1	32	1.102	1.102
2	32		1.115
3	32	1.042	

Graphics were created for the significant two-way interaction effects of A × CS, N × CS and MP × CS (Figure 4.9 – Figure 4.12). Figure 4.9 showed that as MD increased, variance of effect estimates for x_1 decreased for all levels of sample size. When sample size was at 500, 1000, and 5000, variance of effect estimates related to x_1 was the highest at the lowest level of class separation (i.e., MD = 1.0), and when sample size reached 10000, variances were very close to 0 at both MD = 2.0 and MD = 3.5, suggesting that when sample size and class separation were both large, variance of covariate effect estimates related to x_1 was very close to its lower bound of 0.

It was very interesting to observe the interaction effect of class separation and mixing proportion on the variance of covariate effect estimates for x_1 (see Figure 4.10). Variance values almost overlapped around the value of 0 for the two mixing proportion levels at MD = 3.5, indicating that when class separation was very large, covariate effect estimates for x_1 were very stable for both latent class proportion levels. Differences in variance between the two mixing proportion levels was observed at MD = 1.0 and MD = 2.0 where the variance is higher for mixing proportion of 30:70 than that of 50:50. It was also noticed that when class separation was at MD = 2.0, variance of effect estimates for x_1 was closer to 0 at mixing proportion of 50:50 than for the mixing proportion of 30:70.

Effects from the two-way interactions of estimation approach and class separation on the variance of covariate effect estimates were displayed in Figure 4.11 and Figure 4.12 for x_1 and x_2 respectively where a similar pattern was observed. As might be expected, for both x_1 and x_2, variance values were always close to 0 for all estimation approaches at MD = 3.5. Variance values were also always close to 0 for all class separation levels when the conventional three-step approach and the PC approach were used, which makes sense considering their comparatively higher percent relative bias values observed in Section 4.2.1. It was also observed that for the one-step ML method, the covariate effect estimates for both x_1 and x_2 were close to 0 for MD = 2.0 and MD = 3.5. For the three-step ML approach, variance of covariate effect estimates at MD = 1.0

was very large, suggesting that parameter estimates related to both x_1 and x_2 had more variability from using the three-step approach when class separation was very poor.

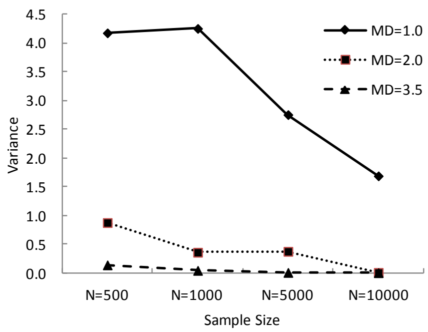


Figure 4.9. N × CS on variance of effect estimates for x_1

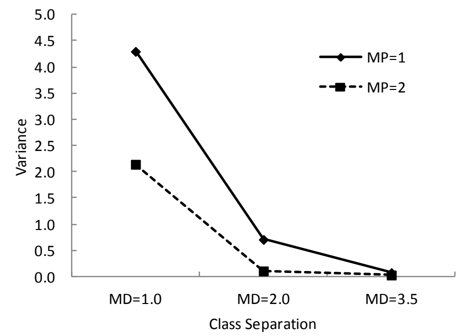


Figure 4.10. CS × MP on variance of effect estimates for x_1

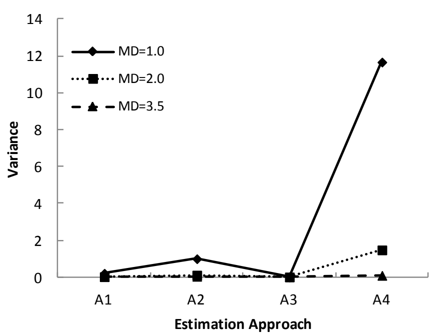


Figure 4.11. A × CS on variance of effect estimates for x_1

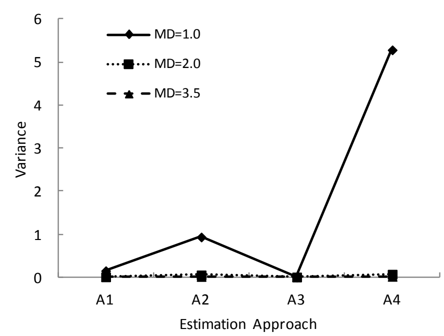


Figure 4.12. A × CS on variance of effect estimates for x_2

Figures 4.13 – 4.16 depicted the three-way interaction effect of A × CS × CE on the variance of parameter estimates related to x_2. Two-way interactions of A × CS were graphed separately for each level of covariate effect. It was observed that for all levels of covariate effect, variances of covariate effect estimates for x_2 were close to 0 for the class separation of MD = 3.5 for all estimation approaches. Also, for all levels of covariate effect, the conventional three-step method and the PC method always showed the lowest variance values across all class separation levels. When class separation was at the lowest considered level of MD = 1.0, both of the one-step ML method and the three-step ML method showed largest variance values for all levels of covariate effect, suggesting again that these two approaches were sensitive to low class separation in terms of variability of covariate effect estimates related to x_2 .

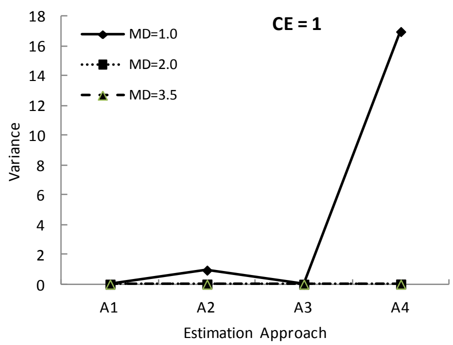


Figure 4.13. A × CS on variance for x_2 at CE=1

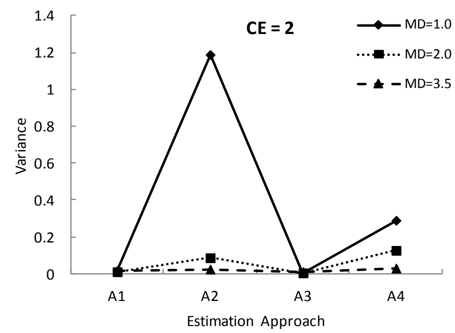


Figure 4.14. A × CS on variance for x_2 at CE=2

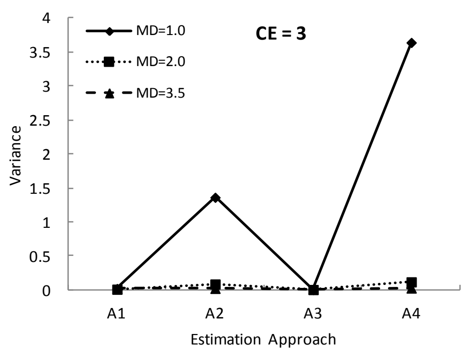


Figure 4.15. A × CS on variance for x_2 at CE=3

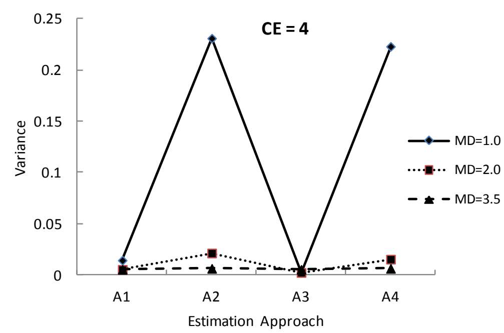


Figure 4.16. A × CS on variance for x_2 at CE=4

4.2.3 Results of standard error efficacy of the covariate effect estimates

4.2.3.1 Descriptive statistics of the standard error efficacy of the covariate effect estimates

Standard error efficacy of the covariate effect estimates, a standard error ratio, was another criterion used to measure the performance of the estimation approaches. As was mentioned earlier, a standard error efficacy value greater than 1 indicates that the standard errors are overestimated, implying increase of making a Type II error. On the other than, if an efficacy value is less than 1, the standard errors are underestimated, suggesting chance of committing a Type I error. Therefore, an efficacy value close to 1 is desired which suggests that the estimated standard errors computed based on an approach provide accurate estimates of the population standard errors.

Tables 4.24 – 4.27 below showed descriptive statistics in standard error efficacy for the covariate effect estimates. It was observed that for all combined levels of sample size, mixing proportion, and covariate effect, standard error efficacy values related to both x_1 and x_2 were closest to 1 for all estimation approaches when class separation was at the highest considered level of MD = 3.5. The PC approach always attained efficacy values greater than 1 across all levels of sample size, mixing proportion and class separation, suggesting higher probability of committing Type II errors when the PC approach was used.

Table 4.24

Standard Error Efficacy with Small Covariate Effects for x_1 and x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	0.210	0.898	1.172	0.942	2.735	2.131	0.354	0.193
		MD = 2.0	1.289	1.005	0.940	1.351	2.014	1.650	1.969	1.005
		MD = 3.5	1.085	1.004	1.059	1.026	1.167	1.146	1.085	1.016
	50:50	MD = 1.0	1.324	1.194	1.360	0.964	2.3797	2.238	0.292	0.060
		MD = 2.0	1.225	0.951	1.229	1.064	1.937	1.528	1.259	0.964
		MD = 3.5	1.086	0.949	1.107	0.947	1.208	1.043	1.089	0.954
1000	30:70	MD = 1.0	2.093	1.074	1.667	0.925	2.923	2.264	0.240	0.043
		MD = 2.0	1.171	0.945	1.184	1.040	1.957	1.630	1.356	0.956
		MD = 3.5	1.020	1.005	1.029	1.015	1.147	1.119	1.019	1.005
	50:50	MD = 1.0	2.496	1.188	1.170	1.229	2.715	2.297	0.322	0.699
		MD = 2.0	1.099	0.990	1.076	0.996	1.776	1.575	1.083	0.975
		MD = 3.5	1.061	0.985	1.057	0.978	1.159	1.082	1.062	0.991
5000	30:70	MD = 1.0	1.266	1.086	4.056	3.258	3.044	2.356	0.395	0.522
		MD = 2.0	0.987	1.005	1.009	1.022	1.734	1.669	0.957	0.955
		MD = 3.5	1.019	1.009	1.015	1.006	1.130	1.124	1.022	1.006
	50:50	MD = 1.0	2.486	1.110	2.480	2.193	2.818	2.046	0.396	0.462
		MD = 2.0	0.985	1.011	1.041	1.063	1.681	1.687	1.009	1.023
		MD = 3.5	0.989	1.021	0.986	1.022	1.088	1.120	0.990	1.024
10000	30:70	MD = 1.0	1.069	0.979	1.124	1.332	2.519	2.429	0.377	0.124
		MD = 2.0	0.964	1.016	0.952	1.051	1.629	1.679	0.952	1.017
		MD = 3.5	0.960	1.041	0.960	1.015	1.064	1.137	0.960	1.039
	50:50	MD = 1.0	1.407	1.146	1.148	1.287	2.456	1.717	0.346	1.122
		MD = 2.0	1.012	1.030	1.048	1.052	1.700	1.657	1.008	1.007
		MD = 3.5	1.017	1.026	1.013	1.039	1.114	1.134	1.020	1.036

Table 4.25

Standard Error Efficacy with Large Covariate Effects for x_1 and x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	1.230	0.756	1.218	1.123	2.446	1.415	0.335	0.852
		MD = 2.0	0.914	0.823	1.319	1.085	1.648	1.373	0.757	0.939
		MD = 3.5	0.960	0.996	1.004	0.987	1.203	1.215	0.973	0.990
	50:50	MD = 1.0	1.245	0.793	1.266	1.261	1.979	1.464	0.375	0.862
		MD = 2.0	0.972	0.794	1.141	1.170	1.645	1.159	0.570	1.076
		MD = 3.5	0.986	0.956	1.015	0.978	1.185	1.163	1.012	1.004
1000	30:70	MD = 1.0	1.516	0.663	1.249	1.246	2.409	1.270	0.362	0.642
		MD = 2.0	0.958	0.924	1.164	1.070	1.792	1.434	1.071	0.922
		MD = 3.5	1.031	0.993	1.069	1.026	1.304	1.197	1.083	1.030
	50:50	MD = 1.0	1.087	0.680	1.123	1.234	1.882	1.405	0.432	2.224
		MD = 2.0	0.964	0.882	1.026	1.048	1.630	1.261	0.893	0.814
		MD = 3.5	0.987	0.972	1.009	0.962	1.188	1.172	1.004	0.988
5000	30:70	MD = 1.0	0.921	0.598	1.016	0.984	2.076	1.028	0.339	0.464
		MD = 2.0	0.843	0.951	1.008	0.944	1.665	1.438	0.924	0.844
		MD = 3.5	0.929	0.939	0.957	0.955	1.163	1.162	0.961	0.962
	50:50	MD = 1.0	1.117	0.562	1.045	1.033	1.560	0.886	0.511	0.504
		MD = 2.0	0.992	1.019	0.982	1.062	1.666	1.382	0.876	0.845
		MD = 3.5	1.008	1.032	1.000	1.056	1.208	1.226	1.018	1.071
10000	30:70	MD = 1.0	0.969	0.557	1.018	0.962	1.895	0.964	0.330	2.867
		MD = 2.0	0.865	0.959	0.981	0.961	1.671	1.449	0.899	0.883
		MD = 3.5	0.915	0.982	0.941	0.998	1.144	1.211	0.943	0.992
	50:50	MD = 1.0	1.107	0.621	1.038	1.030	1.709	0.974	0.435	0.365
		MD = 2.0	0.954	0.986	1.012	1.045	1.654	1.391	0.885	0.803
		MD = 3.5	0.997	0.996	0.962	0.994	1.190	1.214	0.978	1.003

Table 4.26

Standard Error Efficacy with Small Covariate Effect for x_1 and Large Covariate Effect for x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	1.513	0.717	1.307	1.103	3.055	1.506	0.343	0.752
		MD = 2.0	1.467	0.920	1.356	1.139	2.478	1.443	0.988	0.937
		MD = 3.5	1.180	0.988	1.181	1.013	1.435	1.262	1.182	0.982
	50:50	MD = 1.0	1.623	0.726	1.190	1.081	3.082	1.376	0.394	1.325
		MD = 2.0	1.328	0.783	1.282	1.045	2.307	1.183	3.342	1.226
		MD = 3.5	1.163	0.965	1.188	0.994	1.360	1.159	1.192	1.013
1000	30:70	MD = 1.0	2.016	0.689	1.165	1.177	3.252	1.394	0.368	0.612
		MD = 2.0	1.286	0.912	1.254	1.084	2.332	1.369	1.343	0.917
		MD = 3.5	1.137	1.019	1.130	1.048	1.351	1.236	1.148	1.037
	50:50	MD = 1.0	1.383	0.703	1.477	1.329	3.114	1.340	0.235	0.074
		MD = 2.0	1.269	0.907	1.234	0.994	2.158	1.279	1.295	0.809
		MD = 3.5	1.217	1.013	1.117	0.962	1.277	1.155	1.112	0.993
5000	30:70	MD = 1.0	1.588	0.674	1.121	1.035	3.870	1.048	1.148	0.794
		MD = 2.0	1.050	0.938	1.081	0.969	1.982	1.388	1.073	0.853
		MD = 3.5	1.068	0.967	1.065	0.958	1.263	1.176	1.088	0.978
	50:50	MD = 1.0	1.843	0.655	1.114	1.061	2.878	0.899	0.477	0.065
		MD = 2.0	0.993	1.010	1.031	1.059	1.767	1.392	1.018	0.879
		MD = 3.5	1.008	1.004	1.022	1.009	1.186	1.191	1.019	1.026
10000	30:70	MD = 1.0	1.199	0.414	1.133	1.004	3.452	0.985	0.507	0.442
		MD = 2.0	0.987	0.938	1.008	0.999	1.835	1.377	0.988	0.884
		MD = 3.5	1.006	1.000	1.004	1.029	1.190	1.217	1.011	1.037
	50:50	MD = 1.0	1.461	0.477	1.087	1.051	3.141	0.662	4.241	0.334
		MD = 2.0	0.988	0.982	1.013	1.016	1.792	1.344	0.995	0.841
		MD = 3.5	1.032	1.015	1.015	1.042	1.188	1.229	1.031	1.052

Table 4.27

Standard Error Efficacy with Large Covariate Effect for x_1 and Small Covariate Effect for x_2

Conditions			Variance of $\hat{\gamma}$							
			Conventional 3-Step		1-Step ML		PC		3-Step ML	
Sample Size	Mixing Proportion	Class Separation	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
500	30:70	MD = 1.0	1.653	0.721	1.397	0.940	1.907	2.151	0.378	0.944
		MD = 2.0	0.850	0.997	1.892	1.040	1.528	1.623	0.473	1.004
		MD = 3.5	0.980	0.989	1.039	0.976	1.210	1.105	1.163	0.992
	50:50	MD = 1.0	0.235	0.788	1.250	1.082	1.714	2.304	0.372	0.864
		MD = 2.0	0.897	1.013	1.423	0.920	1.425	1.614	0.529	0.992
		MD = 3.5	0.966	1.000	0.980	0.987	1.117	1.109	0.964	1.006
1000	30:70	MD = 1.0	0.182	0.638	1.828	1.042	1.989	2.123	0.405	0.952
		MD = 2.0	0.980	1.007	1.920	1.030	1.721	1.653	0.695	0.995
		MD = 3.5	1.041	0.975	1.031	0.968	1.267	1.085	1.046	0.981
	50:50	MD = 1.0	0.863	1.084	1.671	1.281	1.602	2.212	0.320	0.612
		MD = 2.0	0.960	1.032	1.040	1.014	1.489	1.658	1.177	1.005
		MD = 3.5	0.960	0.995	0.972	0.994	1.106	1.121	0.966	1.001
5000	30:70	MD = 1.0	0.861	0.989	2.496	1.036	1.634	2.483	0.294	0.772
		MD = 2.0	1.010	0.993	1.080	1.031	1.663	1.724	0.611	1.018
		MD = 3.5	0.965	0.995	0.981	0.994	1.148	1.113	0.965	0.996
	50:50	MD = 1.0	1.530	1.193	1.476	1.282	1.235	2.411	0.277	0.334
		MD = 2.0	1.001	1.044	1.018	1.042	1.501	1.729	0.828	1.006
		MD = 3.5	1.050	0.983	1.064	1.001	1.209	1.117	1.057	0.985
10000	30:70	MD = 1.0	0.801	0.815	4.595	1.197	1.539	2.247	0.592	0.452
		MD = 2.0	0.854	1.007	0.940	1.058	1.627	1.745	0.793	1.028
		MD = 3.5	0.926	1.032	0.939	1.026	1.133	1.143	0.936	1.039
	50:50	MD = 1.0	0.939	0.824	1.349	1.212	1.176	1.783	0.817	1.032
		MD = 2.0	0.977	1.046	0.964	1.075	1.493	1.724	0.839	1.014
		MD = 3.5	1.025	1.075	1.026	1.095	1.178	1.211	1.035	1.086

4.2.3.2 Repeated measures ANOVA results for the standard error efficacy of the covariate effect estimates

Results of the repeated measures ANOVA for the standard error efficacy of the covariate effect estimates are presented in Table 4.28. Estimation methods and covariate effect both had significant main effect for x_1 and x_2. The effect sizes of estimation approach on standard error efficacy were $\eta^2 = 0.37$ and $\eta^2 = 0.40$ for x_1 and x_2, respectively, and those of covariate effects on standard error efficacy were $\eta^2 = 0.22$ and $\eta^2 = 0.23$ for x_1 and x_2. Class separation had a significant main effect on standard error efficacy only related to x_1. In terms of interaction effects, A × CS, A × CE, and CE × CS all had significant effects on standard error efficacy for both x_1 and x_2. Significant two-way interaction effects of N × CS and N × CE on standard error efficacy were found related to x_1 ($\eta^2 = 0.08$) and x_2 ($\eta^2 = 0.11$), respectively. Significant three-way interaction effect on standard error efficacy was found for A × CE × CS for both x_1 and x_2, and for N × CE × CS related only to x_2.

Table 4.28

ANOVA Results of Manipulated Factors on the Standard Error Efficacy

Source	x_1			x_2		
	F Value	*p*-value	η^2	F Value	*p*-	η^2
Within-Replications Effects[1]						
A	118.967	.000	0.37	113.842	.000	0.40
A × CS	35.041	.000	0.22	21.983	.000	0.16
A × CE	6.639	.000	0.06	7.147	.000	0.08
A × CE × CS	4.081	.000	0.08	5.697	.000	0.12
Between-Replications Effects						
CS	32.998	.000	0.26			
CE	18.732	.000	0.22	19.276	.000	0.23
N × CE				3.057	.021	0.11
N × CS	3.343	.022	0.08			
CE × CS	3.743	.014	0.09	10.250	.000	0.24
N × CE × CS				2.533	.028	0.18

Note: [1] the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary.
A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion.

Tukey's HSD procedure was used for comparing pairs of means for the main effects of class separation and covariate effects. The means for groups in homogeneous subsets were displayed in Tables 4.29 – 4.31. Table 4.29 showed that when sample size increased, standard error efficacy of the covariate effect estimates for x_1 became closer to the desired value of 1 from 1.431 at MD = 1.0 to 1.071 at MD = 3.5, and the decrease in standard error efficacy values for x_1 was significant for all possible pairs of levels of class separation. Table 4.30 and Table 4.31 showed that when both x_1 and x_2 had large effect, standard error efficacy related to x_1 obtained values close to 1.

Table 4.29

Pairwise Comparisons among Levels of CS for Standard Error Efficacy for x_1

CS	Sample Size	Subset		
		1	2	3
1	32			1.431
2	32		1.254	
3	32	1.071		

Table 4.30

Pairwise Comparisons among Levels of CE for Standard Error Efficacy for x_1

CE	Sample Size	Subset	
		1	2
1	24		1.315
2	24	1.114	
3	24		1.444
4	24	1.136	

Table 4.31

Pairwise Comparisons among Levels of CE for Standard Error Efficacy for x_2

CE	Sample Size	Subset	
		1	2
1	24		1.155
2	24	1.036	
3	24	.991	
4	24		1.155

Two-way interaction effects of A × CS, A × CE, and CE × CS on standard error efficacy were graphed for x_1 and x_2 together to easily compare how these effects impact standard error efficacy of covariate effect estimates related to x_1 and x_2. Figures 4.17 and 4.18 showed that for all estimation approaches, standard error efficacy for both covariates were the closest to the desired value of 1 at MD = 3.5 and the furthest from 1 for MD = 1.0. For all levels of class separation, the PC approach showed efficacy values furthest from 1 compared with the other three approaches. Figure 4.19 and Figure 4.20 showed that for all levels of covariate effect standard error efficacy from using the PC approach were always larger than and much further away from 1 when compared with the values obtained with other three approaches. In terms of two-way interaction effect of CE × CS, Figure 4.21 and Figure 4.22 showed that when class separation was at its largest considered level of MD = 3.5, standard error efficacy values were close to the desired value of 1 for all levels of covariate effect. When class separation was at its lowest considered level of MD = 1.0, standard error efficacy values were found to be further away from 1 at covariate effect levels of 1, and 3 where either both covariates had small effect or x_1 had small effect and x_2 had large effect size.

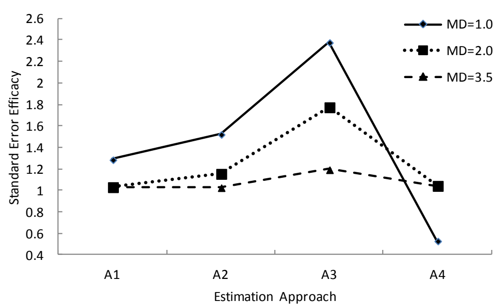


Figure 4.17. A × CS on standard error efficacy for x_1

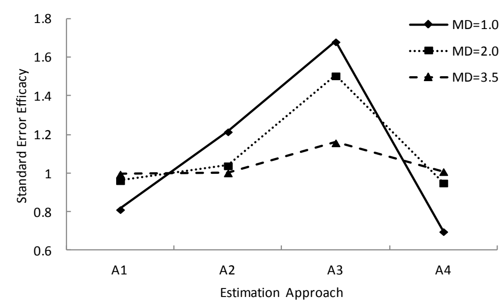


Figure 4.18. A × CS on standard error efficacy for x_2

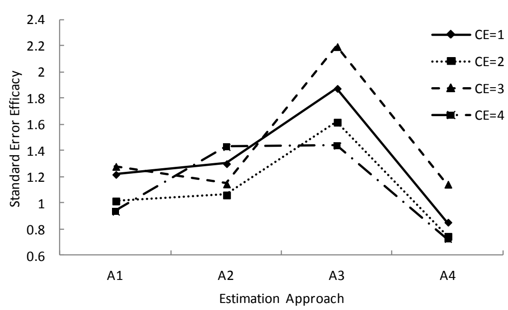


Figure 4.19. A × CE on standard error efficacy for x_1

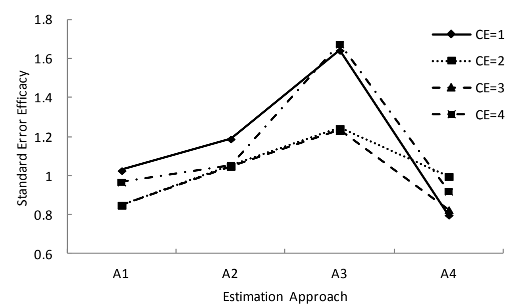


Figure 4.20. A × CE on standard error efficacy for x_2

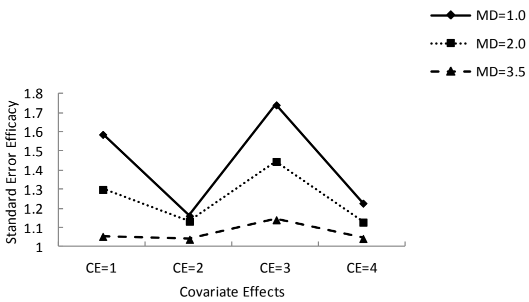


Figure 4.21. CE × CS on standard error efficacy for x_1

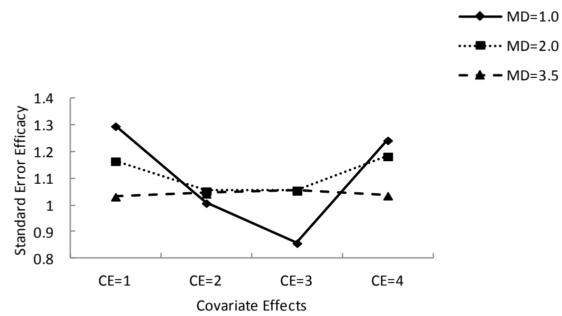


Figure 4.22. CS × CE on standard error efficacy for x_2

Figure 4.23 and Figure 4.24 followed showed the interaction effects of N × CS related to x_1 and N × CE related to x_2, respectively. For the interactions of sample size and class separation, Figure 4.23 showed that at MD = 2.0 and MD = 3.5, standard error efficacy values related to x_1 decreased and tended to approach 1 when sample size was increased. Further when class separation was at MD = 3.5 and sample size was 10000, standard error efficacy was the closest to 1. Two-way interaction effect from sample size and covariate effects on standard error efficacy of covariate effect estimates for x_2 looked more complicated (see Figure 4.24). Standard error efficacy values seemed to be closest to 1 at CE = 3 when sample size was 1000. At CE = 4 standard error efficacy values were very close to each other between levels of sample size and away from 1.

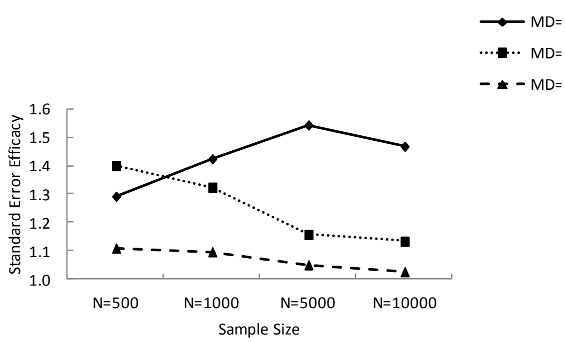


Figure 4.23. N×CS on standard error efficacy for x_1

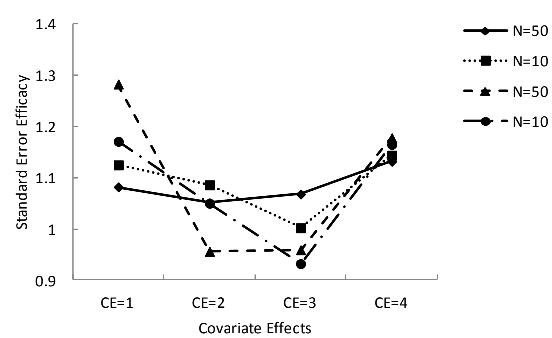


Figure 4.24. N×CE on standard error efficacy for x_2

Figures 4.25 – 4.28 below showed the three-way interaction effect of N × CE × CS on standard error efficacy of the covariate effect estimates related to x_2 by levels of covariate effect. When class separation was at MD = 3.5, standard error efficacy values were close to 1 for all sample sizes at all levels of covariate effect. At MD = 1.0, the efficacy values tended to change or fluctuate a lot among levels of sample size. Efficacy values were comparatively stable among levels of sample size for all levels of covariate effect at MD = 2.0 and MD = 3.5, suggesting that when class separation was large, standard error efficacy of the covariate effect estimates related to continuous variable were close to 1.

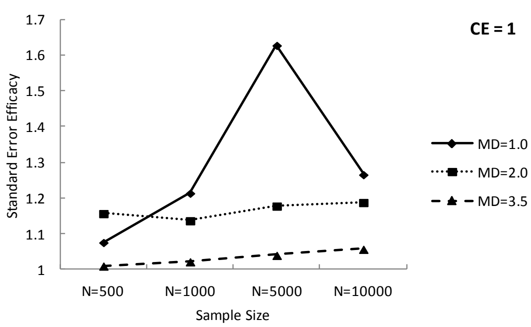


Figure 4.25. N×CS on standard error efficacy for x_2 at CE=1

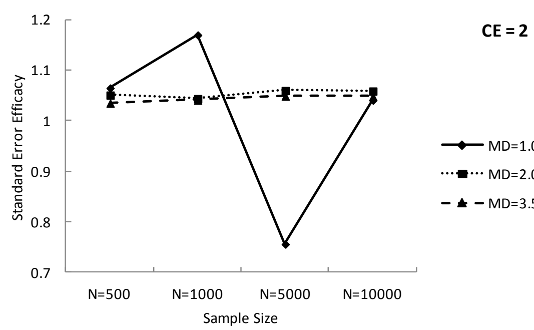


Figure 4.26. N×CS on standard error efficacy for x_2 at CE=2

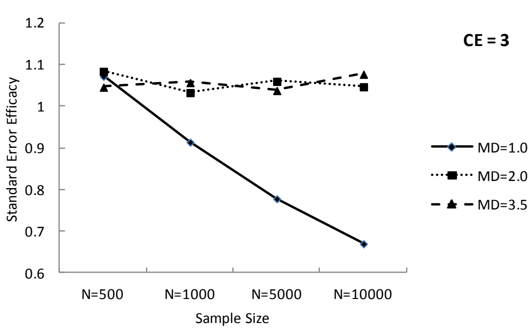


Figure 4.27. N×CS on standard error efficacy for x_2 at CE=3

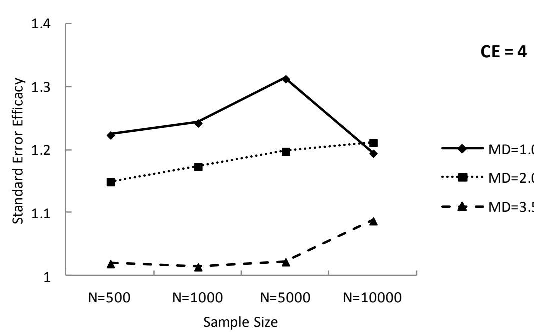


Figure 4.28. N×CS on standard error efficacy for x_2 at CE=4

The three-way interaction effect of A × CE × CS on standard error efficacy related to x_1 and x_2 was presented in four pairs of graphs by the levels of covariate effects (Figures 4.29 – 4.36). It was observed that at CE = 1 and CE = 4, for both x_1 and x_2, all estimation approaches lead to standard error efficacy values close to 1 when class separation was as large as MD = 3.5. When covariate effect was at CE = 1 (where both of the covariate effects had small effect size) and class separation was at MD = 2.0, the conventional three-step procedure, the one-step ML procedure and the three-step ML procedure had standard error efficacy values closer to 1 than the PC procedure. Similarly, when covariate effect was at CE = 2, all the estimation approaches except for the PC method had standard error efficacy values close to 1 at MD = 3.5. When CE = 3, efficacy values for the two covariates were very similar to each other between MD = 2.0 and MD = 3.5 for all the estimation approaches except for the PC procedure.

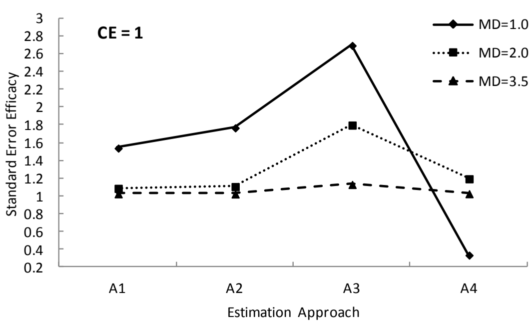


Figure 4.29. A×CS on standard error efficacy for x_1 at CE=1

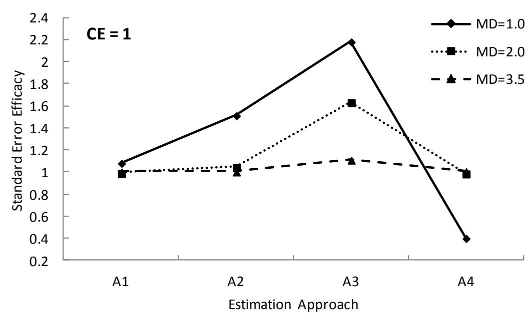


Figure 4.30. A×CS on standard error efficacy for x_2 at CE=1

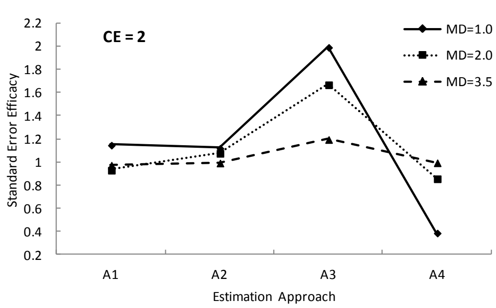


Figure 4.31. A×CS on standard error efficacy for x_1 at CE=2

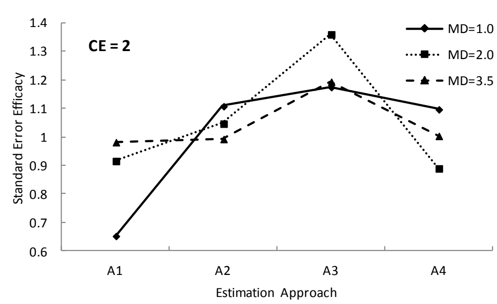


Figure 4.32. A×CS on standard error efficacy for x_2 at CE=2

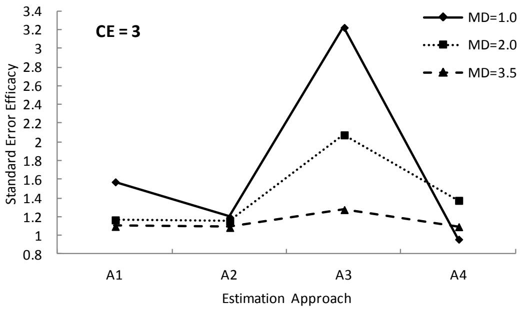


Figure 4.33. A×CS on standard error efficacy for x_1 at CE=3

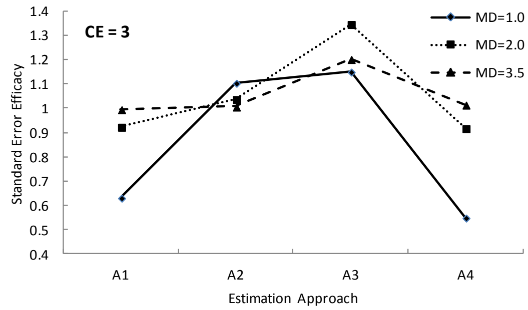


Figure 4.34. A×CS on standard error efficacy for x_2 at CE=3

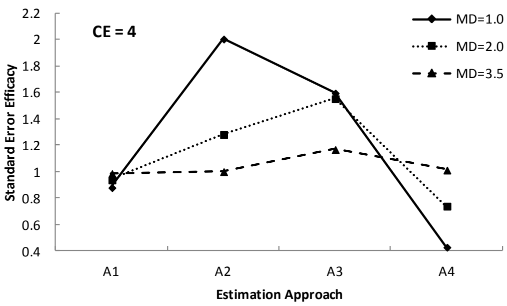


Figure 4.35. A×CS on standard error efficacy for x_1 at CE=4

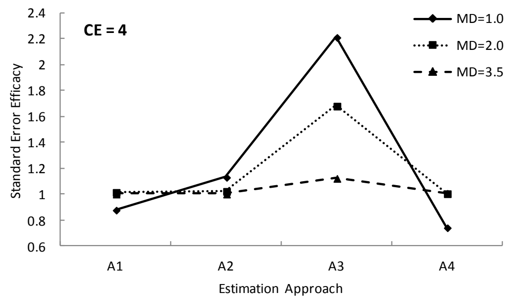


Figure 4.36. A×CS on standard error efficacy for x_2 at CE=4

4.3 Results of Simulation II

As was mentioned earlier in Chapter 3, Simulation II examined how well the conventional three-step approach, the one-step ML approach and the new three-step ML approach performed in terms of covariate effect estimation. Since there were more covariates in the data and these covariates entered different parts of the growth mixture model, it would be interesting to investigate the performance of the three estimation approaches under different model specifications. Therefore, data were analyzed with two models, namely, a misspecified model, and the correctly specified model which was used for data generation. The misspecified model used in the current research in fact was an underspecified model which incorporated only one covariate (linked to the latent class part of the model) and did not include the two covariates supposed to go into the growth part of the model.

In this section, results of Simulation II are reported in the same way as was done in Simulation I. Both descriptive statistics of outcome measures as well as results of several repeated measures ANOVAs were presented using tables or graphs. Specifically, descriptive statistics are provided in two tables separately for the misspecified model and for the correctly specified model in terms of percent relative bias, variance and standard error efficacy of the covariate effect estimates from using the three different estimation approaches. Results of the repeated measures ANOVA were presented separately for the three outcome measures for each of the two models.

4.3.1 Descriptive statistics of the outcome measures for the two models

Table 4.32 and Table 4.33 below showed the descriptive statistics of the three outcome measures by the manipulated conditions for the misspecified model and the correctly specified model respectively. An examination of the percent relative bias values from the three estimation approaches suggested that for both models that were estimated, the conventional three-step approach produced the most biased parameter estimates and consistently underestimated the covariate effect across all conditions. Values of the percent relative bias presented in Table 4.32 showed that for the misspecified model the new three-step ML approach was closer to the desired value of 0 than the one-step ML approach which was in turn closer to 0 than the conventional three-step approach, suggesting that the three-step ML approach resulted in less biased parameter estimates than the other two approaches and the conventional approach always had the poorest covariate effect estimates. It was very interesting to notice in Table 4.33 that for the correctly specified model, percent relative bias values were closer to 0 for the new three-step ML approach than for the one-step approach at small covariate effect whereas percent relative bias values were closer to 0 for the one-step approach than for the three-step ML approach at large covariate effect across all the other simulated conditions, indicating that the three-step ML approach resulted in less biased parameter estimates than the one-step approach when covariate effect from the dichotomous variable was small and that the one-step approach performed better than the three-step approach when covariate effect from the dichotomous variable was large.

In terms of variance of covariate effect estimates, Table 4.32 and Table 4.33 both showed that for both the misspecified model and the correctly specified model, the conventional three-step approach always resulted in the smallest variances and the new three-step ML approach had the largest variances across all condition levels. It was also observed that when covariate effect increased, variance of covariate effect estimates at the same combined conditions of sample size, mixing proportion and class separation increased across all three estimation approaches. Table 4.32 also showed that for the misspecified model, when sample size increased at each combined level of mixing proportion, class separation and covariate effect, variance values decreased for all three estimation approaches. This same consistency was also observed under the one-step ML approach for the correctly specified model.

In terms of the standard error efficacy, when compared with the desired value of 1, for some cells the standard error efficacy values showed very large deviation from 1 when using the three-step ML approach. For example, for the misspecified model at the sample size of 500, mixing proportion of 30:70, class separation at MD = 2.0, and large covariate effect, standard error efficacy of the covariate effect estimation for the three-step approach was 61.580. When sample size increased to 1000 under the same combined condition, standard error efficacy value was 43.718, and when sample size was further increased to 10000, standard error efficacy value was as high as 86.585. Another observation was that for the three-step ML approach, the correctly specified model resulted in standard error efficacy values closer to 1 than the misspecified model at MP = 30:70 and MD = 2.0 at both covariate effect levels across all levels of sample size.

Table 4.32

Outcome Measures for Model 1

Conditions				Relative Bias (%)			Variance			Standard Error Efficacy		
N	MP	CS	CE	A1	A2	A4	A1	A2	A4	A1	A2	A4
500	30:70	MD=2.0	1	-45.3	-33.8	28.2	0.022	0.041	1.223	1.656	1.982	8.988
			2	-61.5	-30.7	-5.8	0.071	0.272	4.478	1.003	4.302	**61.580**
		MD=3.5	1	-14.3	-13.2	-5.5	0.037	0.043	0.045	1.166	1.169	1.167
			2	-14.8	-2.1	2.0	0.105	0.159	0.858	0.987	1.004	0.663
	50:50	MD=2.0	1	-50.9	-34.2	-23.4	0.018	0.050	0.055	1.499	1.771	1.527
			2	-53.7	-20.2	-17.3	0.071	0.152	1.521	0.858	4.088	1.093
		MD=3.5	1	-16.6	-14.9	-9.0	0.031	0.038	0.039	1.117	1.085	1.115
			2	-14.7	-5.8	-3.9	0.062	0.080	0.093	0.944	0.937	0.944
1000	30:70	MD=2.0	1	-59.9	-56.6	-20.8	0.011	0.019	0.101	1.589	1.761	6.843
			2	-60.4	-30.2	-2.2	0.029	0.163	1.849	1.071	1.903	**43.718**
		MD=3.5	1	-23.1	-23.0	-15.4	0.022	0.024	0.027	1.062	1.092	1.061
			2	-16.9	-5.0	-1.0	0.049	0.068	0.113	0.991	1.025	0.994
	50:50	MD=2.0	1	-65.2	-60	-45.9	0.008	0.014	0.019	1.614	1.724	1.605
			2	-53.2	-21.8	-14.0	0.035	0.073	0.964	0.813	1.121	0.907
		MD=3.5	1	-24.4	-24.6	-17.7	0.016	0.018	0.019	1.096	1.093	1.099
			2	-14.7	-6.0	-3.9	0.028	0.034	0.043	0.987	1.003	0.969
5000	30:70	MD=2.0	1	-85.1	-84.9	-60.7	0.005	0.003	0.007	1.027	1.997	1.728
			2	-57.4	-28.5	27.8	0.007	0.024	1.149	1.005	1.147	3.546
		MD=3.5	1	-28.4	-28.3	-21.3	0.005	0.006	0.006	0.999	1.003	0.998
			2	-17.4	-6.1	-2.2	0.009	0.014	0.020	1.001	1.000	0.963
	50:50	MD=2.0	1	-86.5	-86.2	-72.4	0.004	0.003	0.004	1.035	1.742	1.584
			2	-56.1	-21.7	-12.8	0.211	0.014	0.051	0.147	0.985	0.705
		MD=3.5	1	-26.8	-26.8	-20.3	0.004	0.004	0.004	1.048	1.036	1.047
			2	-14.8	-6.0	-4.1	0.006	0.007	0.008	0.980	0.982	0.972
10000	30:70	MD=2.0	1	-90.3	-90	-69.2	0.004	0.001	0.003	0.858	2.068	1.704
			2	-72.7	-40.9	-27.8	0.548	0.012	0.966	0.081	1.166	**86.585**
		MD=3.5	1	-29.4	-29.8	-22.4	0.003	0.003	0.003	0.990	0.979	0.992
			2	-17.8	-6.6	-2.9	0.005	0.007	0.010	0.959	0.949	0.936
	50:50	MD=2.0	1	-89.2	-91.6	-78.0	0.002	0.001	0.002	0.965	2.115	1.661
			2	-55.9	-22.0	-13.3	0.198	0.007	0.023	0.107	0.963	0.721
		MD=3.5	1	-28.0	-27.9	-21.6	0.002	0.002	0.002	0.996	0.996	0.996
			2	-14.7	-5.9	-4.0	0.003	0.004	0.004	0.962	0.972	0.965

Note: A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion; M1: misspecified model; M2: correctly specified model. The bolded numbers are the numbers discussed in Section 4.3.1.

Table 4.33

Outcome Measures for Model 2

Conditions				Relative Bias (%)			Variance			Standard Error Efficacy		
N	MP	CS	CE	A1	A2	A4	A1	A2	A4	A1	A2	A4
500	30:7	MD=2.	1	-63.2	57.8	51.6	0.014	0.055	0.142	1.93	1.664	1.802
			2	-91.8	-5.7	10.2	0.022	0.352	3.949	1.52	5.746	5.640
		MD=3.	1	-50.3	17.0	16.5	0.021	0.042	0.045	1.53	1.169	1.156
			2	-82.6	4.9	9.0	0.042	0.170	0.818	1.09	1.061	1.033
	50:5	MD=2.	1	-67.2	67.1	52.4	0.010	0.062	0.070	1.99	1.308	1.687
			2	-92.5	0.8	-7.6	0.014	0.139	0.964	1.72	1.290	0.655
		MD=3.	1	-54.3	13.6	6.2	0.016	0.038	0.044	1.56	1.083	1.007
			2	-83.3	-0.3	-2.9	0.026	0.079	0.213	1.22	0.944	0.601
1000	30:7	MD=2.	1	-74.4	57.2	56.0	0.006	0.027	0.740	2.02	1.502	0.464
			2	-93.8	-5.1	15.3	0.008	0.154	1.718	1.73	1.272	7.645
		MD=3.	1	-60.4	7.0	6.8	0.013	0.023	0.025	1.36	1.096	1.076
			2	-84.0	2.0	4.7	0.022	0.071	0.113	1.06	1.016	0.957
	50:5	MD=2.	1	-76.9	45.3	33	0.005	0.021	0.030	2.03	1.385	1.205
			2	-93.6	0.0	-4.2	0.007	0.084	0.205	1.62	1.432	1.032
		MD=3.	1	-65.5	3.9	1.9	0.009	0.018	0.020	1.44	1.102	1.044
			2	-83.9	-0.1	-0.2	0.013	0.033	0.046	1.24	1.015	0.904
5000	30:7	MD=2.	1	-87.7	37.3	31.2	0.001	0.004	0.010	1.92	1.656	1.379
			2	-95.1	-2.8	52.1	0.003	0.028	1.931	1.40	1.133	2.075
		MD=3.	1	-75.6	1.6	1.5	0.003	0.005	0.006	1.19	1.012	1.010
			2	-83.6	0.9	3.2	0.005	0.014	0.019	0.98	1.010	0.978
	50:5	MD=2.	1	-87.0	24.0	17.7	0.001	0.003	0.004	1.77	1.486	1.406
			2	-94.1	0.6	-1.1	0.002	0.016	0.037	1.28	0.986	0.772
		MD=3.	1	-74.9	1.5	1.3	0.003	0.004	0.004	1.13	1.050	1.049
			2	-82.3	-0.3	-0.3	0.003	0.007	0.007	1.13	1.000	1.031
1000	30:7	MD=2.	1	-90.7	30.5	26.1	0.001	0.002	0.004	1.91	1.694	1.516
			2	-94.9	-2.1	65.8	0.001	0.013	1.115	1.33	1.158	2.954
		MD=3.	1	-77.4	0.5	0.2	0.002	0.003	0.003	1.08	1.010	1.016
			2	-83.6	0.4	2.7	0.003	0.007	0.009	0.98	0.972	0.972
	50:5	MD=2.	1	-89.3	18.7	12.8	0.001	0.001	0.002	1.67	1.621	1.426
			2	-93.9	0.3	-1.4	0.001	0.007	0.017	1.34	1.002	0.781
		MD=3.	1	-77.4	0.3	0.2	0.002	0.002	0.002	1.00	1.012	1.020
			2	-81.8	-0.1	0.1	0.001	0.004	0.004	1.19	0.980	0.985

Note: A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion; M1: misspecified model; M2: correctly specified model.

4.3.2 Results of repeated measures ANOVA for Simulation II

4.3.2.1 Repeated measures ANOVA results for the percent relative bias

A repeated measures ANOVA was used for both the misspecified model and the correctly specified model to examine the impact of factors and/or combination of factors on percent relative bias for the covariate effect estimate under the three estimation approaches. Percent relative bias was modeled as functions of the manipulated factors of estimation approach, sample size, latent class mixing proportion, class separation and covariate effect size. Estimation approach was used as the only within-replications factor in both the misspecified model and the correctly specified model. Results for up to 3-way interactions as well as the main effects were reported in Table 4.34 only if they were both statistically significant (*p*-value $\leq$ .05) and had medium effect size of $\eta^2 \geq 0.06$ (Cohen, 1988). The sphericity assumption was checked and the Huynh-Feldt correction was considered to adjust the degrees of freedom when the sphericity assumption was not adequately satisfied. In addition, post hoc tests were performed for the significant main effect with at least three groups.

The ANOVA results presented in Table 4.34 showed that except for mixing proportion which had a significant main effect on percent relative bias for only the correctly specified model ($\hat{\eta}^2 = 0.06$), all the other factors had significant effects on percent relative bias of covariate effect estimates for both of the misspecified model and the correctly specified model. More two-way interaction effects were identified significant for the misspecified model than for the correctly specified model. For example, significant two-way interaction effects for the misspecified model included A × MP ($\hat{\eta}^2 = 0.06$), A × CS ($\hat{\eta}^2 = 0.19$), A × CE ($\hat{\eta}^2 = 0.08$), N × CE ($\hat{\eta}^2 = 0.11$), and those for the correctly specified model included only N × CE ($\hat{\eta}^2 = 0.18$) and CS × CE ($\hat{\eta}^2 = 0.11$). No significant three-way interaction effect was identified in the ANOVA analysis for either of the estimated models.

Table 4.34

ANOVA Results of Manipulated Factors on Percent Relative Bias for x_1

Source	M1			M2		
	F Value	*p*-value	η^2	F Value	*p*-value	η^2
Within-Replications Effects[1]						
A	94.049	.000	0.50	1806.783	.000	0.93
A×MP	11.036	.010	0.06			
A×CS	36.332	.000	0.19			
A×CE	15.336	.004	0.08			
Between-Replications Effects						
N	19.311	.018	0.08	10.934	.040	0.10
MP				19.100	.022	0.06
CE	170.328	.001	0.25	113.132	.002	0.33
CS	281.290	.000	0.41	40.278	.008	0.12
N×CE	25.285	.012	0.11	20.388	.017	0.18
CS×CE				38.235	.009	0.11

Note: [1] the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary.
A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion; M1: misspecified model; M2: correctly specified model.

Tukey's HSD procedure was used for comparing pairs of means for the main effects of sample size for both of the models. The means for groups in homogeneous subsets were displayed below in Table 4.35 which showed that when sample size increased, percent relative bias of the covariate effect estimates for x_1 tended to depart from the desired value of 0 for both of the models. Percent relative bias values decreased from -19.1 to -36.3 for the misspecified model and decreased from -9.8 to -22.3 for the correctly specified model when the sample size increased from 500 to 10000, with significant change in relative bias found for both models when sample size increased from 500 to 1000 and for the correctly specified model when sample size increased also from 5000 to 10000, which suggested that covariate effect estimation was more accurate when sample size was small. It was also observed that relative bias values were closer to 0 for the correctly specified model than for the misspecified model at each level of sample size, suggesting that at the same sample size level, covariate effect estimates were less biased for the correctly specified model than for the misspecified model.

Table 4.35

Pairwise Comparisons among Levels of N for Percent Relative Bias for the Two Models

N	Sample Size	Subset			
		M1		M2	
		1	2	1	2
1	8		-19.1		-9.8
2	8	-27.8	-27.8	-16.0	-16.0
3	8	-34.5		-21.3	-21.3
4	8	-36.3		-22.3	

Significant two-way interaction effects were examined using graphs presented in *Figures* 4.37 – 4.42. Interaction effect on percent relative bias between estimation approach and mixing proportion for the misspecified model was depicted in Figure 4.37 where the three-step ML approach always showed relative bias values closer to 0 at both mixing proportion levels. When latent class mixing proportion was at 30:70 (i.e., MP = 1), the three-step ML approach resulted in less biased covariate effect estimates than at the mixing proportion level of 50:50, although no obvious difference in relative bias was observed for either the conventional three-step approach or the one-step approach between levels of mixing proportion. Figure 4.38 depicted the interaction effect of estimation approach and class separation on percent relative bias for the misspecified model. It may be observed that all three estimation approaches had relative bias values closer to 0 at class separation of MD = 3.5 than at MD = 2.0, suggesting that when class separation increased, covariate effect estimates tended to be more accurate for any of these estimation approaches. Figure 4.38 also showed that for the misspecified model, percent relative bias values from using the three-step ML approach were lower than the other two approaches at each class separation levels, and that the conventional three-step approach always resulted in values wither larger distance from 0 than either of the other two methods. In terms of the interaction effect between estimation approach and covariate effect for the misspecified model, it was observed in Figure 4.39 that for all estimation approaches examined, relative bias values were closer to 0 when covariate effect size was large. In addition, the one-step approach lead to the least biased covariate effect estimates at CE = 2 whereas the three-step ML approach performed best in covariate effect estimation at CE =1. The two-way interaction effect of sample size and covariate effect on both models were displayed in Figures 4.41 and 4.42. It looked like percent relative bias values were closer to 0 at CE = 2 than at CE = 1 at each sample size level for the misspecified model but closer to 0 at CE = 1 than at CE = 2 for the correctly specified model. Also, percent relative bias magnitudes were relatively stable across all sample size levels for both of the models at CE = 2, indicating sample size did not have much influence on parameter estimates when covariate effect was large. Figure 4.40 showed the interaction effect between class separation and covariate effect for the correctly specified model. Obviously, relative bias values were closer to 0 at CE = 1 at each class separation level and, when covariate effect was large at CE = 2, percent relative bias values were very similar between class separation levels. In fact, both Figures 4.40 and 4.42 showed that for the correctly specified model percent relative bias values were closer to 0 when covariate effect was small either across class separation levels or across levels of sample size.

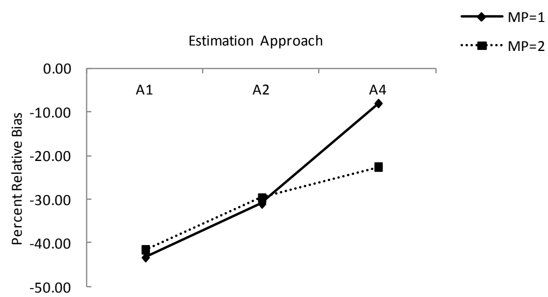


Figure 4.37. A×MP on percent relative bias for Model 1

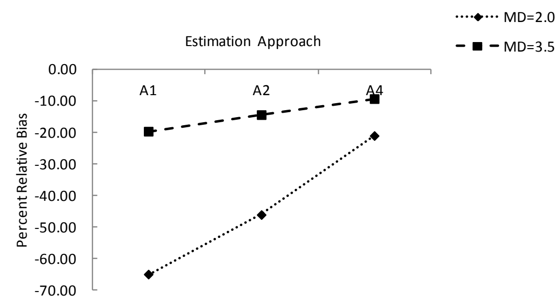


Figure 4.38. A×CS on percent relative bias for Model 1

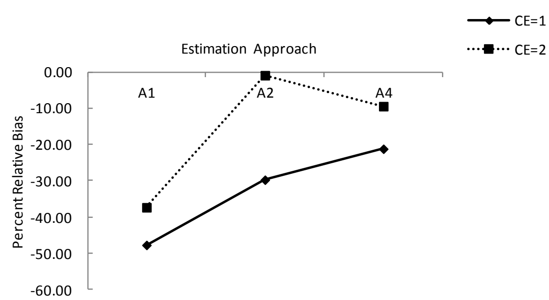


Figure 4.39. A×CE on percent relative bias for Model 1

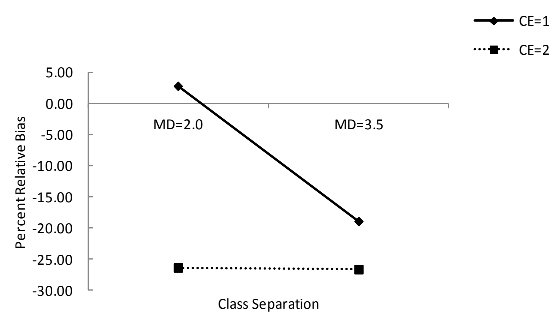


Figure 4.40. CS×CE on percent relative bias for Model 2

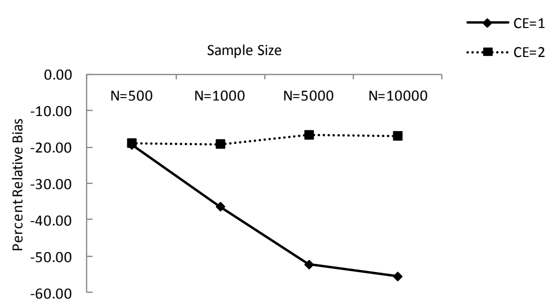


Figure 4.41. N×CE on percent relative bias for Model 1

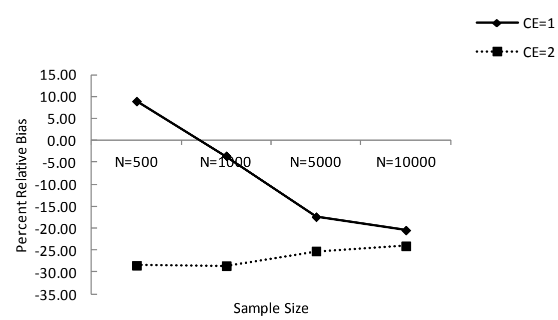


Figure 4.42. N×CE on percent relative bias for Model 2

4.3.2.2 Repeated measures ANOVA results for the variance of the covariate effect estimates

Following the same criteria used before for identifying significant factors and/or combination of factors in repeated measures ANOVA analysis, all manipulated factors showed significant main effects on variances for both the misspecified model and the correctly specified model (see Table 4.36). While moderate two-way interaction effects for MP × CS ($\hat{\eta}^2 = 0.08$), MP × CE ($\hat{\eta}^2 = 0.08$), and CS × CE ($\hat{\eta}^2 = 0.09$) were found only for the correctly specified model, significant two-way interaction effects were found for A × N, A × MP, A × CS, A × CE, and N × CE for both models. Significant three-way interaction effects were identified for A × N × CE and A × CS × CE for both of the two models. In addition, A × N × CS showed a significant three-way interaction effect ($\hat{\eta}^2 = 0.09$) only for the misspecified model while the three-way interactions of A × MP × CS ($\hat{\eta}^2 = 0.07$) and A × MP × CE ($\hat{\eta}^2 = 0.07$) were found significant only for the correctly specified model.

Table 4.36

ANOVA Results of Manipulated Factors on the Variance of Covariate Effect Estimates

Source	M1			M2		
	F Value	*p*-value	η^2	F Value	*p*-value	η^2
Within-Replications Effects[1]						
A	401.489	<.000	0.16	67.347	<.000	0.17
A×N	114.277	<.000	0.14	8.273	.011	0.06
A×MP	160.672	<.000	0.07	38.957	<.000	0.10
A×CS	298.451	<.000	0.12	45.035	<.000	0.12
A×CE	246.873	<.000	0.10	46.842	<.000	0.12
A×N×CE	48.796	<.000	0.06	8.354	.010	0.06
A×CS×CE	170.571	<.000	0.07	30.242	.001	0.08
A×N×CS	70.053	<.000	0.09			
A×MP×CS				28.886	.001	0.07
A×MP×CE				26.233	.001	0.07
Between-Replications Effects						
N	281.726	<.000	0.17	12.799	.032	0.11
MP	375.695	<.000	0.08	39.275	.008	0.12
CS	771.055	<.000	0.16	43.041	.007	0.13
CE	813.944	<.000	0.17	52.386	.005	0.15
N×CE	98.751	.002	0.06	9.703	.047	0.09
MP×CS				25.979	.015	0.08
MP×CE				26.814	.014	0.08
CS×CE				29.377	.012	0.09

Note: [1] the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary. A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion; M1: misspecified model; M2: correctly specified model.

Results of pairwise comparisons were only conducted for the main effect of sample size which has four levels. Pairs of means for sample size for both the misspecified model and the correctly specified model were compared using Tukey's HSD procedure and the results were presented in Table 4.37. The means for groups in homogeneous subsets suggested that for both models when sample size increased, variance of covariate effect estimates increased.

Table 4.37

Pairwise Comparisons among Levels of N for Variance of Covariate Effect Estimates for the Two Models

N	Sample Size	Subset				
		M1			M2	
		1	2	3	1	2
1	8			0.399		0.306
2	8		0.156		0.142	0.142
3	8	0.076			0.088	
4	8	0.066			0.050	

Figures 4.43 – 4.47 on the next page displayed all the significant two-way interaction effects for the misspecified model. The patterns for these effects were easy to follow. For the interaction effect of sample size and covariate effect (see Figure 4.43), variance values decreased at both covariate effect levels when sample size increases, and variance was larger at CE = 2 than at CE = 1 for all sample size levels. For the interaction effect of estimation method and sample size, Figure 4.44 showed that variances of covariate effect estimates tended to decrease for all estimation approaches when sample size increased. The decrease in variance between sample size levels was more obvious for the three-step ML approach than for the other approaches, and variance values seemed close between the conventional method and the one-step method at each sample size level. When sample size was at 10000, variance values for all the three approaches were close to each other. Figure 4.45 – Figure 4.47 showed how the interaction effects between estimation approach and mixing proportion, class separation or covariate effect impacted variances of covariate effect estimates. Similar patterns may be noticed in these graphs

where low variance values were found for all estimation approaches at MP = 2, MD = 3.5, or CE = 1. With the new three-step approach, variance values differed greatly between levels of mixing proportion, class separation and covariate effect. In addition, the conventional three-step approach and the one-step approach had close variance values at and between levels of MP, CS and CE.

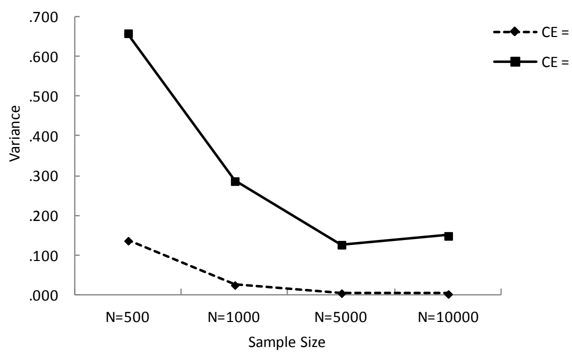


Figure 4.43. N×CE on variance for M1

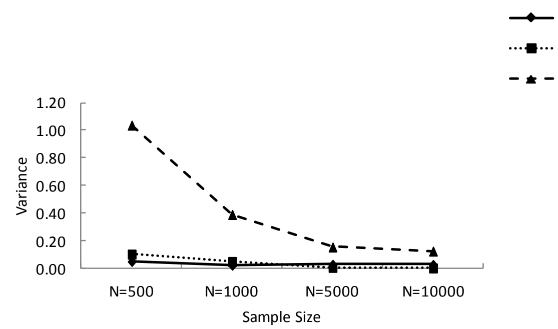


Figure 4.44. A×N on variance for M1

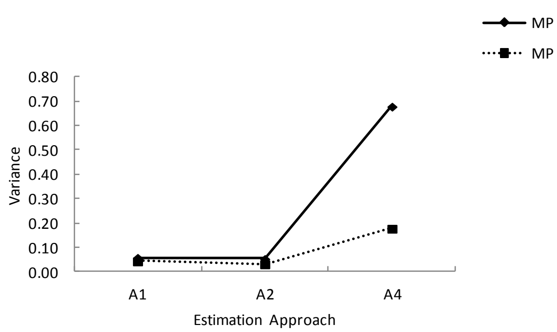


Figure 4.45. A×MP on variance for M1

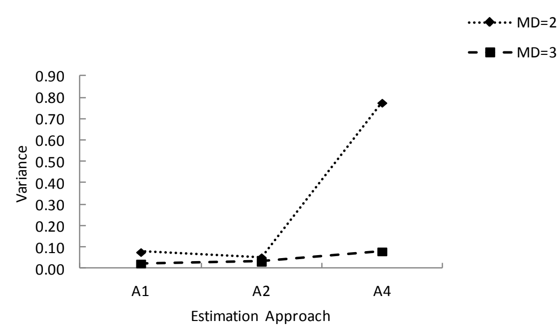


Figure 4.46. A×CS on variance for M1

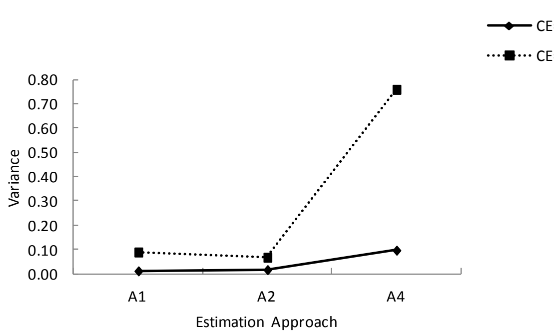


Figure 4.47. A×CE on variance for Model 1

For the significant two-way interaction effect found for the correctly specified model, graphs were also created and displayed on the next page in Figures 4.48 – 4.55. For the interaction effect of sample size and covariate effect (see Figure 4.48), the observation was a little different from Figure 4.43 in that the decrease of variance values corresponding to the increase of sample size was not as obvious as that was observed for the misspecified model. However, the same interaction effect for both models did show that variance values were large at CE = 2 for all sample size levels. Figure 4.49 and Figure 4.50 were for the interaction effects of MP × CS and MP × CE respectively. Variance values were small at MD = 3.5 and at CE = 1 for both mixing proportion levels. Also, at these two factor levels, variance values seemed close between levels of mixing proportion. The interaction effect of class separation and covariate effect (Figure 4.51) showed that variance values were low at MD = 3.5 for both covariate effect levels, and a large discrepancy in variance between class separation levels was found at CE = 2. Figures 4.52 – 4.55 showed how the same within- and between-replication interaction effects examined in the misspecified model affected variances of covariate effect estimates under the correctly specified model, and a comparison between these figures and the figures for the misspecified model suggested that overall all these four interaction effects impacted variances in the same ways no matter which of the two models was considered.

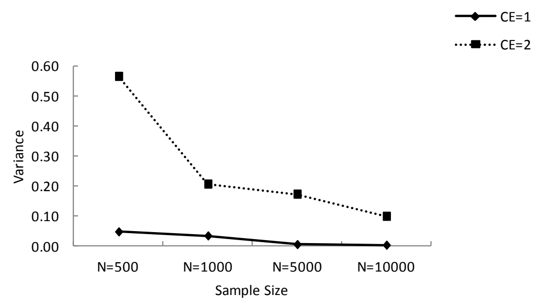


Figure 4.48. N×CE on variance for M2

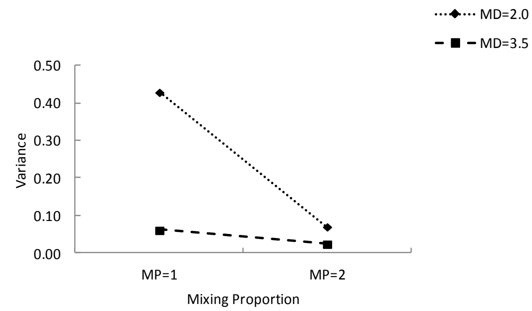


Figure 4.49. MP×CS on variance for M2

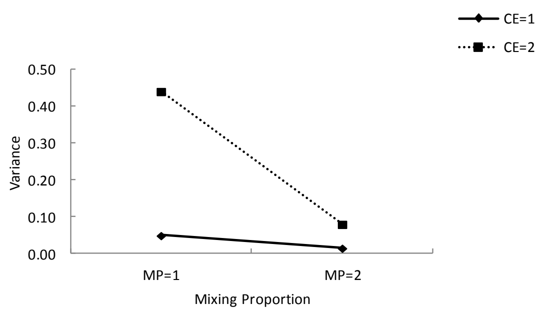


Figure 4.50. MP×CE on variance for M2

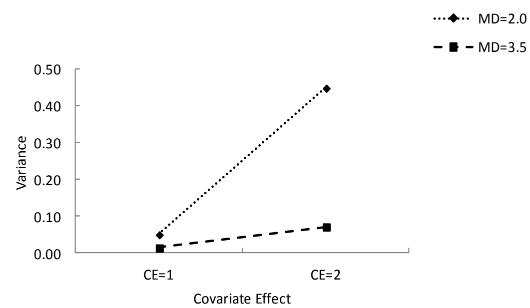


Figure 4.51. CS×CE on variance for M2

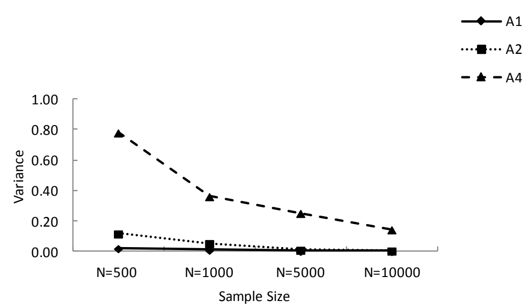


Figure 4.52. A×N on variance for M2

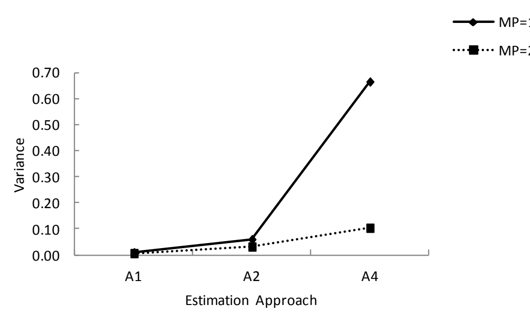


Figure 4.53. A×MP on variance for M2

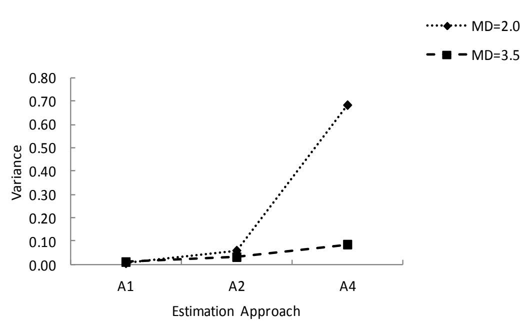


Figure 4.54. A×CS on variance for M2

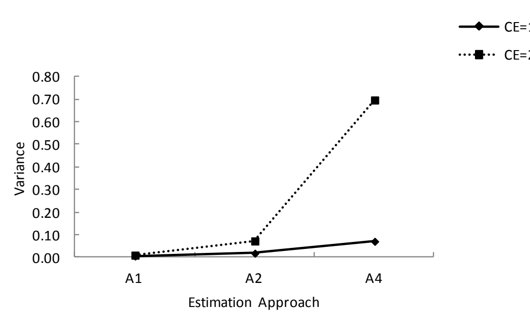


Figure 4.55. A×CE on variance for M2

Significant three-way interaction effects for the misspecified model were displayed in Figures 4.56 – 4.61, and significant three-way interaction effects for the correctly specified model were displayed in Figures 4.62 – 4.69. For the interaction effect of A × N × CS for the misspecified model, Figures 4.56 and 4.57 showed that variance values decreased when sample size increased for all estimation methods at both class separation levels and that this decrease was most obvious for the three-step ML approach. It was also observed that the conventional approach and the one-step approach were close in variance of parameter estimates at each level of sample size. Significant three-way interaction effects identified for both models were: A × N × CE and A × CS × CE. Figures 4.58 and 4.59 showed the interaction effect of estimation approach and sample size at each covariate effect level for the misspecified model. Similar graphics were also created for the correctly specified model displayed in Figures 4.62 and 4.63. The common observation from these two pairs of plots was that variances decreased when sample size increased for all estimation methods at both covariate effect levels. In terms of the interaction effect of A × CS × CE, it was observed that at both levels of covariate effect variances were low at MD = 3.5 for all estimation methods. Also, the conventional approach and the one-step approach had close variance values at each level of class separation, and for the new three-step approach, variances were much smaller at MD = 3.5 at both levels of covariate effect.

Three-way interaction effects identified for the correctly specified model were shown in Figures 4.66 – 4.69. It was observed that variances were low for all estimation methods at MP = 2 for both class separation levels and covariate effect levels.

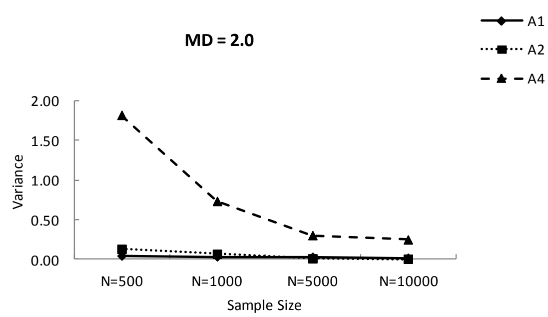


Figure 4.56. A×N on variance at MD=2.0 for M1

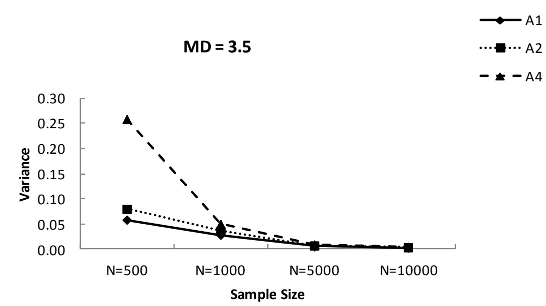


Figure 4.57. A×N on variance at MD=3.5 for M1

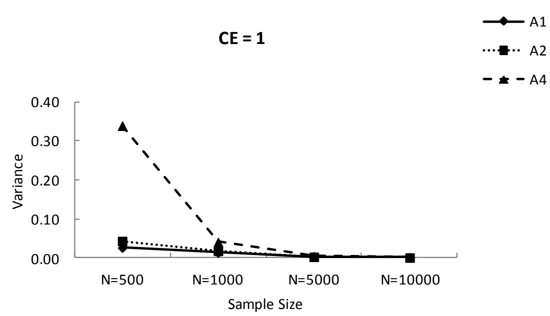


Figure 4.58. A×N on variance at CE=1 for M1

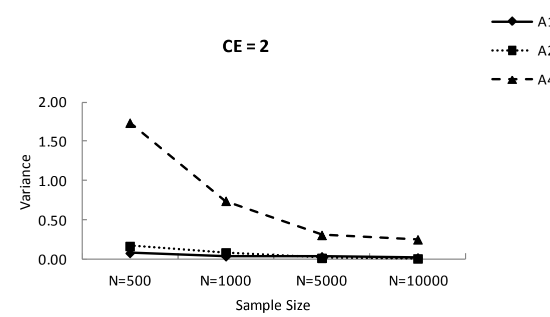


Figure 4.59. A×N on variance at CE=2 for M1

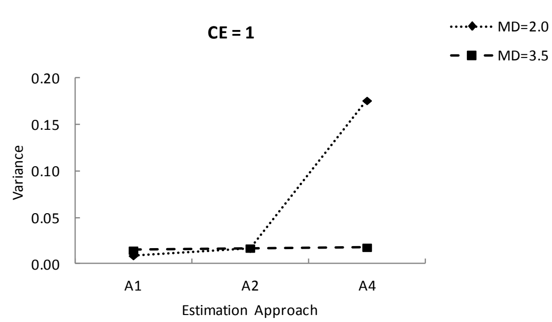


Figure 4.60. A×CS on variance at CE=1 for M1

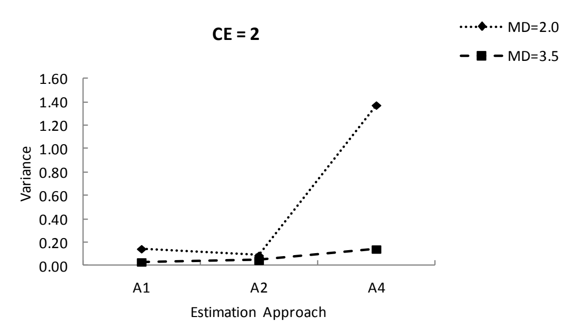
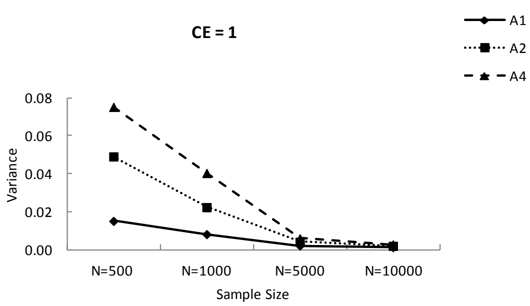


Figure 4.61. A×CS on variance at CE=2 for M1



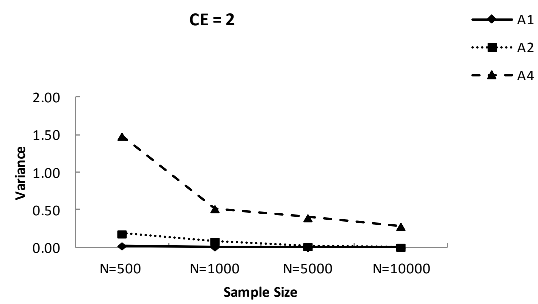
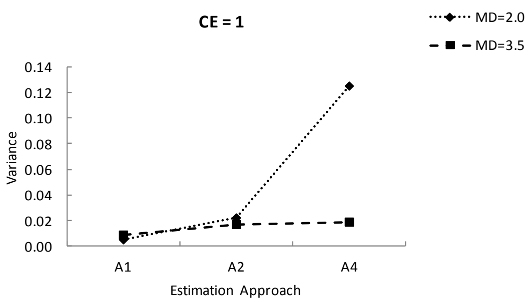


Figure 4.62. A×N on variance at CE=1 for M2

Figure 4.63. A×N on variance at CE=2 for M2



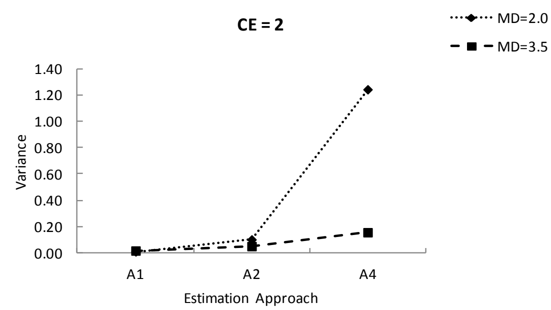
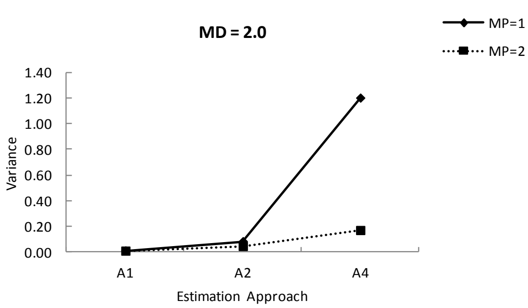


Figure 4.64. A×CS on variance at CE=1 for M2

Figure 4.65. A×CS on variance at CE=2 for M2



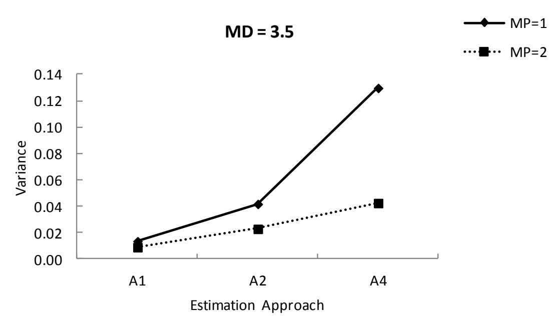
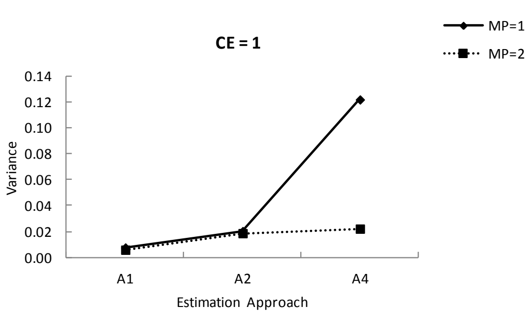


Figure 4.66. A×MP on variance at M =2.0 for M2

Figure 4.67. A×MP on variance at M =3.5 for M2



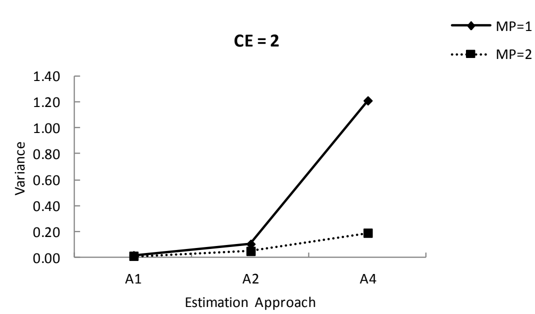


Figure 4.68. A×MP on variance at CE=1 for M2

Figure 4.69. A×MP on variance at CE=2 for M2

4.3.2.3 Repeated measures ANOVA results for the standard error efficacy of the covariate effect estimates

Table 4.38 below provided a listing of the ANOVA results of manipulated factors on the standard error efficacy of the covariate effect estimates related to x_1. The identified significant factors and combination of the factors were shown for both of the two models, and it was observed that none of the factors or combined factors was recognized significant for the correctly specified model. For the misspecified model, interaction effects of A × MP, A × CS, A × CE, A × MP × CS, A × MP × CE, and A × CS × CE are reported because they were both statistically significant (p-value $\leq$ .05) and had an effect size of $\eta^2 \geq 0.06$.

Table 4.38

ANOVA Results of Manipulated Factors on the Standard Error Efficacy for x_1

Source	M1			M2		
	F Value	*p*-value	η^2	F Value	*p*-value	η^2
Within-Replications Effects[1]						
A	8.433	.018	0.10			
A×MP	8.560	.017	0.10			
A×CS	8.504	.018	0.10			
A×CE	6.376	.033	0.08			
A×MP×CS	8.628	.017	0.10			
A×MP×CE	6.564	.031	0.08			
A×CS×CE	6.441	.032	0.08			
Between-Replications Effects						
CS	10.440	.048	0.13			

Note: [1] the Huynh-Feldt correction was used to adjust the degrees of freedom if necessary.
A: covariate estimation approach; CS: class separation; CE: covariate effect; N: sample size; MP: latent class mixing proportion; M1: misspecified model; M2: correctly specified model.

Six interaction effects were graphed for the misspecified model and displayed in Figures 4.70 – 4.78. Figures 4.70 – 4.72 showed the two-way interaction effects for A × MP, A × CS and A × CE, respectively. These three interaction effects resulted in graphs that looked similar to each other although they involved different between-replications factors. It was observed that standard error efficacy values were close to 1 for MP = 2, MD = 3.5, and CE = 1 for all estimation approaches. Standard error efficacy values were close to 1 for the conventional approach and the one-step approach for all levels of mixing proportion, class separation, and covariate effect, and standard error efficacy values departed substantially from 1 at MP = 1, MD = 2.0, and CE = 2 for the three-step ML approach.

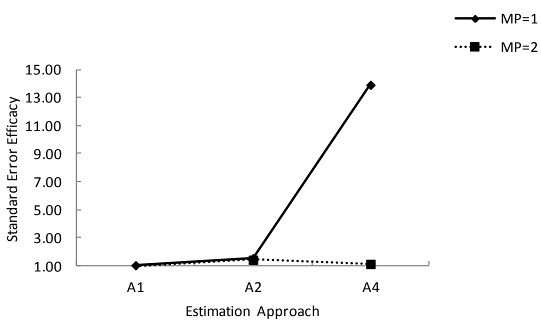


Figure 4.70. A×MP on standard error efficacy for M1

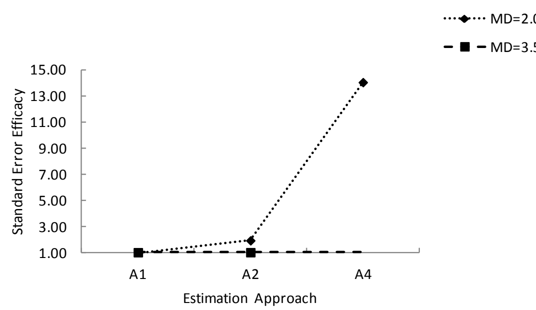


Figure 4.71. A×CS on standard error efficacy for M1

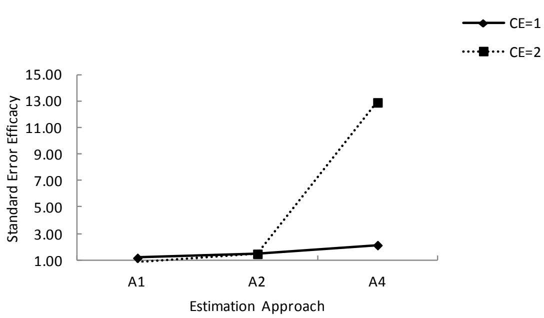


Figure 4.72. A×CE on standard error efficacy for M1

Three-way interaction effects are presented graphically in Figures 4.73 – 4.78 for the misspecified model. For the interaction effect of A × MP × CS, the two-way interactions of A × MP were plotted for each CS level. Similarly, for the interaction effect of A × MP × CE, interaction effect of A × MP were plotted for each CE level, and A × CE were plotted for each CS level for the three-way interaction effect of A × CS × CE. Similar to what was observed for the two-way interaction effects, the three-way interaction effects examined the two-way interaction effects for a third factor condition. It was observed that standard error efficacy values were close to 1 for all the estimation approaches at MP = 2 for both levels of class separation and for both levels of covariate effect. Efficacy values for the three-step ML approach at MP = 1 were much higher than 1, suggesting more chances of making Type II errors with the three-step ML approach when mixing proportion was at 30:70. However, standard error efficacy values were close to 1 for all three estimation approaches at CE = 1 for both class separation levels. It was observed again that standard error efficacy values were further away from the desired value of 1 for the three-step ML approach at MP = 1 across levels of class separation and covariate effect, and at large covariate effect across levels of class separation.

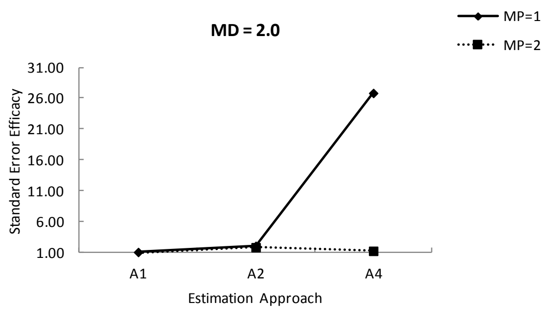


Figure 4.73. A×MP on standard error efficacy at MD=2.0 for M1

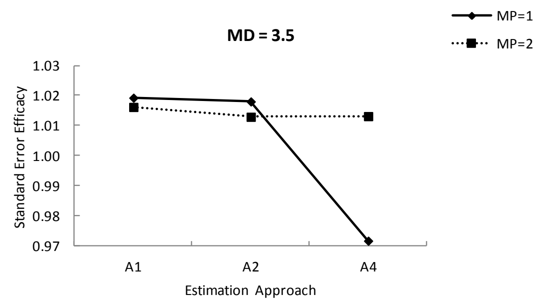


Figure 4.74. A×MP on standard error efficacy at MD=3.5 for M1

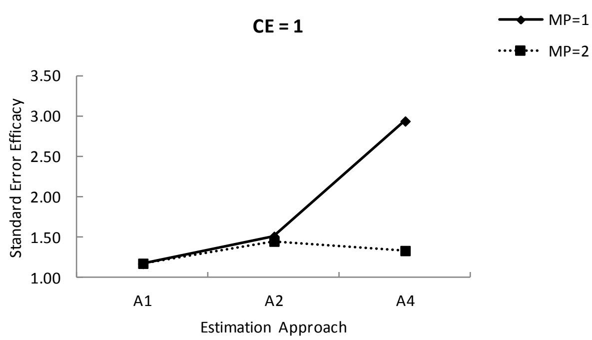


Figure 4.75. A×MP on standard error efficacy at CE=1 for M1

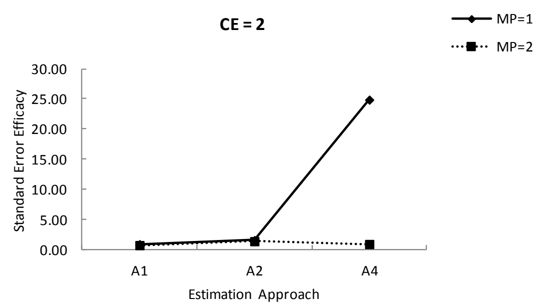


Figure 4.76. A×MP on standard error efficacy at CE=2 for M1

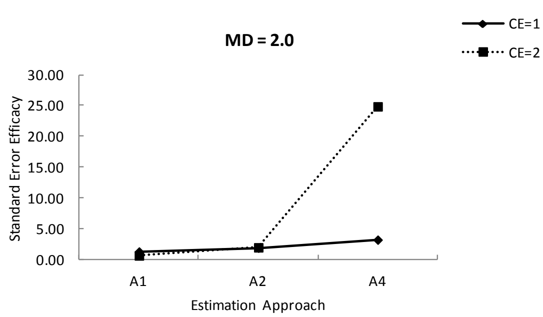


Figure 4.77. A×CE on standard error efficacy at CS=1 for M1

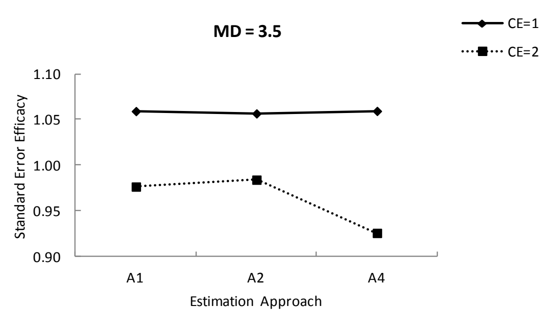


Figure 4.78. A×CE on standard error efficacy at CS=2 for M1

Chapter 5: Discussion

5.1 Discussion of the Simulation Results

The focus of the current research was on evaluating the performance of various methods for estimating covariate effects on the latent class membership using Monte Carlo simulations. The procedures that were compared were the conventional three-step approach, the one-step ML approach, the PC approach, and the new three-step ML approach for Simulation I. The PC approach was not included in Simulation II because of its poor performance observed from Simulation I. Although the two Monte Carlo simulations both examined how well different estimation approaches performed in terms of estimating covariate effects on the latent class membership, they differed mainly in how the data were generated and what models were used for the data analyses. Specifically, Simulation I examined the performance of four estimation methods under the correctly specified measurement model (i.e., the unconditional GMM) where two covariates related to latent class membership were included in the analysis. Simulation II examined the performance of three estimation methods under both the correctly specified model as well as a misspecified model. For the correctly specified model, one covariate entered the latent class part of the model whereas the other two covariates were incorporated in the growth part of the model as they were generated, which made the model more complex compared with the model used in Simulation I. In terms of the misspecified model, data were generated under a correctly specified measurement model but fit with a model where only the covariate linked to latent class membership entered the analysis. This type of misspecification corresponds to real data analytic scenarios in which practitioners may not know whether the GMM should have covariates or not, or what covariates should be included in the model. This misspecified model was considered because we wanted to see how well estimation methods performed for a misspecified model under the simulated conditions. Therefore, in Simulation II we were looking at how well the investigated estimation approaches performed under the manipulated conditions when models were becoming more complicated and when the model was misspecified.

5.1.1 Convergence rate

Although convergence was not the focus of this research per se, it was still useful to get an idea of how well the estimation approaches under investigation performed in terms of converging to a consistent, local solution. Since non-convergence or multiple local maxima are common problems with using EM algorithm for fitting finite mixture models, the choice of an estimation method with fewer convergence issues might be the first concern before researchers start any applied study using mixture models.

Convergence rate results from Simulation I suggested that when class separation was very large, all estimation approaches had 100% convergence rates at each simulated condition. Convergence rates for the PC method and the three-step ML method were above 95% across all conditions. When class separation was as large as MD = 2.0 convergence rates for all estimation methods were high at or above 99% when sample size was 5000 and 10000. Compared with the other three methods, the one-step approach was more sensitive to class separation, sample size, and covariate effects. For example, low convergence rates were observed for the one-step method when class separation was very poor at MD = 1.0 and sample size was small when both of the covariates had small effects or the continuous covariate has small effect. However, convergence rates improved greatly for the one-step approach under the worst conditions of low class separation and low covariate effects when sample size increased to 5000, suggesting that convergence problem for the one-step approach could be mitigated with large sample size (e.g., 5000) under the worst condition where the continuous variable had small effect and class separation was poor.

Results from Simulation II showed that the convergence rates for the three estimation methods were higher at any manipulated condition when the model used for the analysis was correctly specified than when the model was misspecified, suggesting that model specification might be an important factor to impact model convergence. For the correctly specified model, the convergence rates were improved for all three estimation methods across levels of covariate effects and mixing proportion when sample size and class separation increased. When class separation was very large at MD = 3.5, convergence rates were high and very close between the two models for both the one-step and the new three-step ML approaches across other conditions. Also, when the model was correctly specified, the convergence rates were generally higher for the one-step method than for the three-step ML method when covariate effect was large, which suggested again that the one-step approach is sensitive to covariate effect size. In fact, the convergence rates were much higher for the one-step approach than for the other approaches with both models when class separation was large.

5.1.2 Different approaches for covariate effect estimation

Performance of various approaches for estimating covariate effects on the latent class membership was investigated under the same manipulated factors for the two simulations. However, due to the poor performance of the PC method found in

Simulation I and the very low convergence rates from low class separation found in a pilot study for Simulation II, only three estimation methods and two levels of class separation were considered for the second simulation. In addition, because only one covariate was related to latent class membership, only two levels of covariate effect were manipulated in Simulation II. Also, for the first simulation, performance of the estimation approaches was investigated using the true model, and for the second study, both a misspecified model and the true model were fit to the same data. For the two true models used, we wanted to see how well the selected approaches performed in covariate effect estimation in terms of recovery and standard error efficacy under similar manipulated conditions when a model got more complicated. In other words, we were really interested in knowing how the estimation methods interacted with the other manipulated factors under each model in terms of covariate effect estimate accuracy. For the misspecified model used in Simulation II, we wanted to see how well the selected methods performed in covariate effect estimation under the manipulated factors when a simple model was used for data analyses. Therefore, in this research the estimation approaches were examined in terms of covariate effect estimation on the latent class membership under three different models, using both descriptive statistics and repeated ANOVAs. Percent relative bias, variance of covariate effect estimates and standard errors of the covariate effect estimates were used as criteria for evaluating the estimation approaches under investigation.

5.1.2.1 Findings from Simulation I

Results of both the descriptive statistics and the repeated measures ANOVA for Simulation I showed that estimation approach had a large impact on the accuracy of parameter estimates of interest. When class separation was very large, all of the four approaches tended to have less biased parameter estimates at each combined manipulated condition. The PC method and the conventional three-step approach lead to more biased parameter estimates, which was consistent with previous findings by Vermunt (2010). It was also found that covariate estimates related to both the dichotomous and the continuous variables for the PC approach were more biased than for the conventional three-step approach across all combined manipulated conditions. Consistent with the findings of Asparouhov and Muthén (2013), when class separation was very large, the one-step and the three-step ML approaches resulted in very close and more accurate covariate effect estimates across all levels of covariate effects. It was also found that parameter estimate related to the dichotomous covariate was severely affected by poor class separation and small covariate effect related to the dichotomous variable when the three-step ML approach was used. Corresponding to what was found about percent relative bias, the one-step ML approach and the three-step ML method had more variability in covariate effects estimation than the conventional three-step method or the PC method, and that for all covariate effects levels, the conventional three-step method and the PC method always showed the least variability across all class separation levels. In terms of standard error efficacy of the covariate effect estimates, results showed that the efficacy values for both covariates were the closest to 1 when class separation was very large and the furthest from 1 when class separation was poor for all the estimation methods. Standard error efficacy values greater than 1 for the PC method meant more chances of committing Type II errors from using this method. It was also found that when the covariate effects were small for both auxiliary variables, all estimation approaches lead to standard error efficacy values close to 1 when class separation was as large as MD = 3.5. Standard error bias from using either the conventional three-step approach or the new three-step ML approach were close to 1 when class separation was large.

5.1.2.2 Findings from Simulation II

Results of both the descriptive statistics and the repeated measures ANOVA for Simulation II indicated that for both the misspecified and the correctly specified models, the conventional three-step approach not only consistently underestimated the covariate effect of the variable related to the latent class membership but the parameter estimates were the most biased.

For the misspecified model, the three-step ML approach resulted in the least biased parameter estimates. With respect to variances, the values tended to decrease for all estimation approaches when sample size increased, and the decrease in variance values between sample size levels was more obvious for the three-step ML approach than for the other approaches. It was also interesting to find that for both the misspecified model and the correctly specified model, sample size did not have much influence on accuracy of parameter estimates when the covariate effect was large.

In terms of parameter estimation from the misspecified model, results showed that when class separation increased, covariate effect estimates were less biased for the one-step ML and the three-step ML approaches but not for the conventional approach. However, the standard error efficacy values for the conventional approach and the one-step ML approach were much closer to 1 than the three-step ML approach when mixing proportion was 30:70 and the dichotomous covariate had large effect, suggesting more chances of committing Type II errors from using the three-step ML approach under this condition. For the correctly specified model, the three-step ML method had the least biased covariate effect estimates when the dichotomous covariate had small effect whereas the one-step ML approach lead to the least biased parameter estimates when the dichotomous has large effect. For the correctly specified model, parameter estimates from the conventional approach were more biased when covariate effect was large. Variances were small when sample size was large, mixing proportion was 30:70, or class separation was very large for the correctly specified model.

The similarities between the two models make a lot of sense in that for the misspecified model, the covariate effect estimated was related to the variable that entered the latent class part of the model, so the misspecified model was in some sense 'partly' correct with missing only the information from the two other covariates that were supposed to be incorporated into the measurement part of the model. Different from ANOVA results of Simulation I, significant main or interaction effects on variance of covariate effect estimates were not found for the correctly specified model, suggesting that when a model was misspecified, many factors might affect bias and standard error efficacy of the covariate effect, and thus, its corresponding hypothesis test.

5.2 Recommendations for Applied Researchers

Many factors may influence a researcher's choice of a particular estimation method used for his or her study. The following general recommendations are provided with respect to the consideration of model convergence issues and other factors based on the findings from the current research.

In terms of model convergence concerns, it is recommended that class separation be examined before continuing the research. When class separation is large enough, convergence issues may not be a big concern for any of the estimation methods investigated. However, with low class separation, convergence problems might occur with any estimation approach. For example, Simulation I showed that when class separation was poor and a continuous covariate had low effect, convergence rate for the one-step ML approach was low except when sample size was as large as 5000, suggesting that convergence problem for the one-step approach under the worst condition (i.e., the continuous variable has small covariate effect and the class separation is poor) approach could be mitigated when large sample size is used.

Model specification is another very important factor for convergence. The model used for data analysis should take into account the theories in the related field, or when there is not enough theory behind the proposed model, selection of a certain model could be made by comparing fit indexes. In terms of the model selection, in a recent simulation study, Liu and Hancock (2014) proposed the idea of using an unrestricted multivariate normal mixture strategy to assess class enumeration. It was found that the theoretically compelling completely unrestricted multivariate normal mixture model was superior to the linear GMM when the nature of the growth curve was not certain and the sample size was sufficiently large.

In addition to convergence issues, the choice of an estimation method also depends on the accuracy of parameter estimates from using a method, which has to also take into account the characteristics/structure of the data to be analyzed. Based on the findings from this research, the PC method is not recommended, especially when class separation is low. It is also recommended that when covariate effect for a categorical variable is large, the one-step ML method might be a better choice whereas with small covariate effect, the three-step approach performs better in parameter estimation. It should be reminded that large class separation is always important for more accurate parameter estimates when the new three-step ML approach is to be used. It should also be added that in Simulation I parameter estimates related to the dichotomous variable were severely affected by small covariate effect from that variable for the 3-step ML approach when class separation was poor. However, in Simulation II it was found that the three-step ML approach lead to less biased parameter estimates than the one-step approach when covariate effect was small for all levels of class separation. The reason is that in Simulation II, levels of class separation were both large. Therefore, results from both simulations in terms of influence of covariate effect on the three-step ML approach were consistent.

5.3 Implications, Limitations and Future Research

The idea of the current study was stimulated by Vermunt (2010). The study is comprehensive in that instead of looking at only LCA, we examined the approaches for covariate effect estimation under the very complex growth mixture modeling framework. Since as nearly every application in longitudinal research incorporates some covariate information and applied researchers want to know how covariates help explain group membership, it is important that the estimation of the relation between covariates and the latent class membership is accurate when an estimation approach is used.

Like all other studies, the current research has limitations. First, in terms of the experimental design, the manipulated conditions in this research may not generalize to all possible real-life conditions. Second, using only replications that had converged solutions may have impacted generalizability of the inferences drawn from the research, especially when convergence rates were low for some conditions. In addition, increasing the number of iterations in order to obtain the aimed number of converged replications might have made the results inaccurate in the current study.

Third, in terms of the models used in the study, they were not representative of all possible models present in the real world situation in terms of model complexity or model specification. The current research represents a step forward from previous studies by considering more covariates of different types and by considering covariates incorporated into the different parts of a growth mixture model. The situations manipulated in this research were much simpler than real life situations where more often researchers might be faced with a large number of covariates and no information was provided as to which part of the model each covariate is supposed to enter. However, the results could be suggestive of what may happen in these more complex situations. It should be reminded also that the misspecified model selected for Simulation II was in fact an under-specified model which did not include the information from the measurement part, which explains why estimation methods and other manipulated factors interacted similarly between the two models used in Simulation II in terms of impacts on outcome measures. For Future research, more misspecified models should be examined and significant tests should also be conducted to see how estimation methods impact covariate effect estimation under different model specifications.

Fourth, the current study used the converged replications across all estimation approaches for the analyses, which means that the replications that did not converge for any one estimation approach were not used for the other approaches. It would be interesting to examine why some particular replications worked for one estimation approach but not for the others.

Fifth, results from the current research showed that the PC approach performed poorly almost across all simulated conditions. It should be noted, however, that the current research used only the default random draws from M*plus*. It would be interesting to see what the results are like when the number of random draws is increased.

Finally, more estimation should be explored so that the strength of association between covariate effects and growth trajectories could be examined. With that, more interesting research could be done to better understand how covariates are related to different parts of a growth mixture model.

Appendix A

Suppose X_1 and X_2 be random variables with means μ_1 and μ_2, variances σ_1^2, σ_2^2, and covariance σ_{12}, respectively. Let β_1 and β_2 be constants. The algorithm used for growth trajectory related covariate effect control for Simulation II is described below:

Let $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$ is:

$$\mathbf{\Sigma} = E\{(\mathbf{X}-\mathbf{u})(\mathbf{X}-\mathbf{u})'\} = \begin{pmatrix} E(X_1-u_1)^2 & E(X_1-u_1)(X_2-u_2) \\ E(X_1-u_1)(X_2-u_2) & E(X_2-u_2)^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{\beta}'\mathbf{\Sigma}\boldsymbol{\beta} &= (\beta_1 \quad \beta_2)\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\ &= \beta_1^2\sigma_1^2 + \beta_1\beta_2\sigma_{21} + \beta_1\beta_2\sigma_{12} + \beta_2^2\sigma_2^2 \\ &= \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + 2\beta_1\beta_2\sigma_{12} \end{aligned}$$

References

Asparouhov, T., & Muthén, B. O. (2013). *Auxiliary variables in mixture modeling: 3-step approaches using Mplus*. Retrieved from https://www.statmodel.com/download/3stepOct28.pdf

Asparouhov, T., & Muthén, B. O. (2006). *Comparison of estimation methods for complex survey data analysis*. Retrieved from http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_110.pdf

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology.* Advance online publication. Doi: 10.1177/0081175012470644

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375-1386.

Bandalos, D. L., & Leite, W. L. (2013). Use of Monte Carlo Studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 564-666). Greenwich, CT: Information Age Publishing.

Bauer, D. J. (2007). 2004 Cattell award address: Observations on the use of growth mixture

models in psychological research. *Multivariate Behavioral Research*, *42*, 757-786.

Bauer, D. J., & Curran, P. J. (2003). Over-extracting latent trajectory classes: Much ado about nothing? *Psychological Methods*, *8*, 384-393.

Bolck, A., Croon, M. A., & Hagenaars, J. A. P. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3-27.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Brown, E. C. (2003). *Estimates of statistical power and accuracy for latent trajectory class enumeration in the growth mixture model* (Unpublished doctoral dissertation). University of South Florida, Gainesville.

Choi, J., Harring, J. R., & Hancock, G. R. (2009). Latent growth modeling for logistic response functions. *Multivariate Behavioral Research*, *44*, 620-645.

Clark, S., & Muthén, B. (2009*). Relating latent class analysis results to variables not included in the analysis*. Retrieved from http://www.statmodel.com/download/relatinglca.pdf

Clogg, C. C. (1995). Latent class models: Recent developments and prospects for the future. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp.311-352). New York, NY: Plenum.

Codd, C. L., & Cudeck, R. (2014). Nonlinear random-effects mixture models for repeated measures. *Psychometrika*, *79*, 60-83.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colder, C. R., Campbell, R. T., Ruel, E., Richardson, J. L., & Flay, B. R. (2002). A finite mixture model of growth trajectories of adolescent alcohol use: Predictors and consequences. *Journal of Consulting and Clinical Psychology, 70*, 976-985.

Coder, C. R., Mehta, P., Balanda, K, Campbell, R. T., Mayhew, K. P., Stanton, W. R., Pentz, M. A., & Flay B. R. (2001). Identifying trajectories of adolescent smoking: An application of latent growth mixture modeling. *Health Psychology*, *20*, 127-135.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173-178.

Demirtas, H., Amatya, A., & Doganay, B. (2014). BinNor: An R package for concurrent generation of binary and normal data. *Communications in Statistics-Simulation and Computation*, *43*, 569-579.

Dempster, A. P., Laird, N. M, & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*, 1-38.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*, 186-219.

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*, 643-659.

D'Unger A. V., Land, K. C., & McCall, P. L. (2002). Sex differences in age patterns of delinquent/criminal careers: Results from Poisson latent class analyses of the Philadelphia cohort study. *Journal of Quantitative Criminology*, *18*, 349-375.

Ellickson, P. L., Martino, S. C., & Collins, R. L. (2004). Marijuana use from adolescence to young adulthood: Multiple developmental trajectories and their associated outcomes. *Health Psychology*, *23*, 299-307.

Everitt, B. S. (1981). A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, *16*, 171-180.

Feng, X., Shaw, D. S., & Silk, J. S. (2008). Development trajectories of anxiety symptoms among boys across early and middle childhood. *Journal of Abnormal Psychology*, *117*, 32-47.

Fergusson, D. M., & Horwood, L. J. (2002). Male and female offending trajectories. *Development and Psychopathology*, *14*, 159-177.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611-631.

Goodman, L. A. (2007). On the assignment of individuals to latent classes. Sociological Methodology, 37, 1-22.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in M*plus* and SAS. *Structural Equation Modeling*, *16*, 676-701.

Hagenaars, J. A. (1993). *Loglinear models with latent variables.* London: Sage Publications.

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course (2nd ed.)* (pp. 307-340). Charlotte, NC: Information Age Publishing, Inc.

Harring, J. R. (2009). A nonlinear mixed effects model for latent variables. *Journal of Educational and Behavioral Statistics*, *34*, 293-318.

Harring, J. R. (2012). Finite mixtures of nonlinear mixed effects models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences*. Charlotte, NC: Information Age Publishing, Inc.

Heybroek, L. (2011). *Life satisfaction and retirement: A latent growth mixture modeling approach*. Retrieved from http://www.melbourneinstitute.com/downloads/conferences/HILDA_2011/HILDA11_final%20papers/Heybroek.Lachlan_4B_fpaper.pdf

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, *11*, 36-53.

Huang, D. Y. C., Brecht, M. L., Hara, M., & Hser, Y. I. (2010). Influences of a Covariate on Growth Mixture Modeling. *Journal of Drug Issues*, *40*, 173-194.

Huang, D. Y. C., Murphy, D. A., & Hser, Y. I. (2012). Developmental trajectory of sexual risk behaviors from adolescence to young adulthood. *Youth and Society*, *44*, 479-499.

Jo, B., Wang, C. P., & Lalongo, N. S. (2010). Using latent outcome trajectory classes in causal inference. *Stat Interface*, *2*, 403-412.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*, 302-317.

Kohli, N. (2011). *Estimating unknown knots in piecewise linear-linear latent growth mixture models* (Unpublished doctoral dissertation). University of Maryland, College Park.

Li, L., & Hser, Y. (2011). On inclusion of covariates for class enumeration of growth mixture models. *Multivariate Behavioral Research*, *46*, 266-302.

Li, M., Harring, J. R., & Macready, G. B. (2014). Investigating the feasibility of using Mplus in the estimation of growth mixture models. *Journal of Modern Applied Statistical Methods*, *13*, 484-513.

Liu, J. (2012). *A systematic investigation of within-subject and between-subject covariance structures in growth mixture models* (Unpublished doctoral dissertation). University of Maryland, College Park.

Liu, M., & Hancock, G. R. (2014). Unrestricted mixture models for class identification in growth mixture modeling. *Educational and Psychological Measurement*. Advance online publication. Doi:10.1177/0013164413519798

Liu, M., Hancock, G. R., & Harring, J. R. (2011). Using finite mixture modeling to deal with systematic measurement error: A case study. *Journal of Modern Applied Statistical Methods*, *10*, 249-261.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21-39.

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, *14*, 26-47.

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

Magnusson, D. (1966). *Test Theory*. Reading, Mass: Addison Wesley.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, *12*, 49–55.

Mann, H. M. (2009). *Testing for differentially functioning indicators using mixtures of confirmatory factor analysis models.* Unpublished doctoral dissertation,

University of

Maryland, College Park.

Masyn, K., & Brown, E. C. (April, 2001). *Latent class enumeration in general growth mixture modeling*. Presented at the annual meeting of the American Educational Research Association, Seattle, WA.

McArdle, B. H. (1988). The structural relationship: Regression in biology. *Canadian Journal of Zoology*, *66*, 2329–2339.

McCutcheon, A. C. (1987). *Latent class analysis*. Beverly Hills, CA: Sage Publications.

McDermott, S., & Nagin, D. S. (2001). Same or different?: Comparing offender groups and covariates over time. *Sociological Methods and Research*, *29*, *2*82-318.

McIntosh, C. N. (2013). Pitfalls in subgroup analysis based on growth mixture models: A commentary on van Leeuwen et al. *Quality of Life Research*, *22*, 2625-2629.

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

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions*. New York, NY: Wiley.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107-122.

Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-33).Mahwah, NJ: Erlbaum.

Muthén, B. O. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods*, *8*, 369–377.

Muthén, B. O. (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 Publications.

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

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

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371-402.

Muthén, B. O., & Muthén, L. (2000). Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*, 882-891.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the

EM algorithm. *Biometrics*, *55*, 463-469.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, *4*, 139-177.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D. S., Farrington, D. P., & Moffitt, T. E. (1995). Life-course trajectories of different types of offenders. *Criminology*, *33*, 111-139.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology*, *31*, 327-362.

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.

Peteras, H., & Masyn, K (2010). General growth mixture analysis with antecedents and consequences of change. In A. Piquero, & D. Weisburd (Eds.), *Handbook of quantitative criminology*. New York, NY: Springer-Verlag.

Pinquart, M., & Schindler, I. (2007). Changes of life satisfaction in the transition to retirement: A latent-class approach. *Psychology and Aging, 22*, 442-455.

Roeder, K., Lynch, K., & Nagin, D. (1999). Modeling uncertainty in latent class membership: A

case study in criminology. *Journal of the American Statistical Association*, *94*, 766-776.

Rolfe, M. (2010). *Bayesian models for longitudinal data* (Unpublished doctoral dissertation). Queensland University of Technology.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons, Inc.

Simon, M., Ercikan, K., & Rousseau, M.(2012). Introduction. In M. Simon, K. Ercikan, & M. Rousseau (Eds), *Improving large-scale assessment in education: Theory, issues, and practice*. New York, NY: Taylor and Francis Routledge.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In Hancock, G. R. & Samuelsen, K. M. (Eds.), *Advances in latent variable mixture models*. Greenwich, CT: Information Age.

Tolvanen, A. (2008). *Latent growth mixture modeling: A simulation study* (Unpublished doctoral

dissertation). University of Jyvaskyla.

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

Tueller, S. J., & Lubke, G. H. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, *17*, 165-192.

Van der Heijden, P. G. M., Dessens, J., & Böckenholt, U. (1996). Estimating the concomitant variable latent class model with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *21*, 215-229.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450-469.

Wang, C. P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth

mixture models: Examining the impact of preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, *100*, 1054-1076.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, *10*,

635-656.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395-426.

Yang, M., & Dunson, D. B. (2010). Bayesian semiparametric structural equation models with latent variables. *Psychometrika*, *75*, 675-693.