ABSTRACT

Title of Document: AN INVESTIGATION OF GROWTH MIXTURE MODELS WHEN DATA ARE COLLECTED WITH UNEQUAL SELECTION PROBABILITIES: A MONTE CARLO STUDY

Jennifer Hamilton, Doctor of Philosophy, 2009

Directed By: Professor Gregory R. Hancock, Department of Measurement, Statistics & Evaluation

As researchers begin to use Growth Mixture Models (GMM) with data from nationally representative samples, it becomes increasingly critical for researchers to understand the difficulties associated with modeling data that come from complex sample designs. If researchers naively apply GMM to nationally representative data sets without adjusting for the way in which the sample was selected, the resulting parameter estimates, standard errors and tests of significant may not be trustworthy.

Therefore, the objective of the current study was to quantify the accuracy of parameter estimates and class assignment when subjects are sampled with unequal probabilities of selection. To this end, a series of Monte Carlo simulations empirically investigated the ability of GMM to recover known growth parameters of distinct populations when various adjustments are applied to the statistical model. Specifically, the current research compared the performance of GMM that 1) ignores the sample design; 2) accounts for the sample design via weighting; 3) accounts for the sample design via explicitly modeling the

stratification variable; and 4) accounts for the sample design by using weights and modeling the stratification variable.

Results suggested that a model-based approach does not improve the accuracy of parameter estimates when individuals are sampled with disproportionate sampling probabilities. Not only does this method often fail to converge, when it did converge the parameter estimates exhibited an unacceptable amount of bias. The weighted model performed the best out of all of the models tested, but still resulted in parameter estimates with unacceptably high percentages of bias. It is possible that the distributions of the manifest variables overlap too much, and the aggregate distribution may be unimodal, making it potentially difficult to distinguish among the latent classes and thus affecting the accuracy of parameter estimates. In sum, the current research indicates that GMM should not be used when data are sampled with disproportionate probabilities. Researchers should therefore attend to the study design and data collection strategies when considering the use of a Growth Mixture Model in the analysis phase.

AN INVESTIGATION OF GROWTH MIXTURE MODELS WHEN DATA ARE COLLECTED WITH UNEQUAL SELECTION PROBABILITIES: A MONTE CARLO STUDY

by

Jennifer Hamilton

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
2009

Advisory Committee:
Professor Gregory R. Hancock, Chair
Professor Robert G. Croninger
Professor Hong Jiao
Professor Robert W. Lissitz
Professor George B. Macready

DEDICATION

I dedicate this dissertation to my family without whom, this achievement wouldn’t have happened; to Missy and Rich for keeping me laughing and to Mum and Dad for your unconditional love. Mike, you have stood beside me since the moment I met you, and I truly appreciate your unwavering faith in me. I can’t thank you enough. Lastly, this dissertation is dedicated to the memory of Biscuit, whose quiet devotion I will never forget. This accomplishment belongs to all of you.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank the people who have helped me. I must first thank the coolest advisor in the world, Dr. Gregory Hancock; your commitment to your students and your love of statistics has truly been an inspiration to me. My sincere thanks also go to my committee members, Dr. Robert Croninger, Dr. Hong Jiao, Dr. Robert Lissitz, and Dr. George Macready. All of your feedback and insight have improved and polished this document.

Lastly, I would not have gotten this far without the continued support of my employer, Westat, and specifically the support of Babette Gutmann and Alex Ratnofsky. Thank you for your patience and your friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PURPOSE AND RATIONALE

1.1 BACKGROUND

Researchers and practitioners from many disciplines have long been interested in understanding how certain variables change and develop over time. Over the past 15 years several new statistical methods for analyzing longitudinal data have been developed. Latent Growth-curve Modeling (LGM) has emerged from the area of structural equation modeling as a popular and relatively simple technique for modeling change at both the individual and group level. LGM identifies an average trajectory for the sample and estimates variability around this average trajectory. However, an important limitation of LGM is that it assumes that all individuals are drawn from an observed population. In many applied research settings, multiple unobserved populations may exist.

Therefore, researchers have recently expanded LGM to include a categorical latent class variable (Muthén, 2001a, b; Muthén & Shedden, 1999; Nagin, 1999). In contrast to LGM, this new technique, called Growth Mixture Modeling, is designed to identify the growth trajectories of two or more unobserved subpopulations, and to estimate the probability of membership into each trajectory. For example, the application of a Growth Mixture Model to college drinking data identified 5 drinking trajectories (Greenbaum, Del Boca, Darkes, Wang, & Goldmen, 2005). Further, this article identified which drinking trajectories were more likely to later develop into problematic drinking patterns. These five trajectory types were modeled as latent classes because population membership was not directly observed.

Raudenbush (2001) observed that LGM and Growth Mixture Modeling answer different types of research questions. Referring to LGM he stated, “In many studies it is reasonable to assume that all participants are growing according to some common function, but the growth parameters vary in magnitude” (p. 30). LGM thus answers research questions such as, “What is the typical pattern of growth and how does it vary across subjects?” However, if trajectories do not vary around a single mean trajectory, the use of Growth Mixture Models (GMM) are recommended because it identifies multiple distinctive population trajectories and estimates the proportion of cases sharing each trajectory type. GMM thus answer a different kind of research question: “What are the characteristics of multiple unobserved populations, and how does each population change over time?” (Raudenbush, 2001; p.30). For example, using GMM, researchers can examine the impact of a new educational program on students in one latent population compared to its impact on students belonging to a second population (e.g., Kreisman, 2003; Muthén, Khoo, & Francis, 1998). GMM are less frequently used to assign subjects to their most likely latent class, such as assigning students into latent classes in order to provide a specific treatment to members of a high risk group (e.g., Boscardin, 2001; Greenbaum et al., 2005; Neuman et al., 1999; Rasmussen et al., 2002).

Although it is still a fairly new analytical technique, research on GMM has already “left the initial phase of method illustrations and has entered the phase of serious substantive applications” (Muthén, 2003, p. 3). However, as applied researchers begin to use GMM to analyze large, nationally-representative data sets, it is important to note that these data are far more complex than those collected via simple random sampling (SRS). In fact, SRS is rarely used in practice for large-scale surveys (Kalton, 1983). Comprehensive reviews of the special characteristics of typical complex sample designs are provided by Longford (1995) and Lee, Forthofer, and Lorimore (1989). Their treatments cover the topics of clustering, stratification, unequal probabilities of selections, and nonresponse and post stratification adjustments. Much of the data collected and distributed by the federal government, such as the Early Childhood Longitudinal Study and the High School and Beyond Longitudinal Study, use such complex sampling designs.

These large scale data collection efforts typically result in data that are clustered at different levels. Because traditional estimation of standard errors assume the correlation of errors across individuals is zero, a researcher using clustered data may underestimate the standard errors. Such underestimation would subsequently result in inflated Type I error rates (Kish & Frankel, 1974). In recent years there has been numerous research studies documenting the effects of such clustering on parameter estimates (e.g., Allua, Stapleton, & Beretvas, 2008; Muthén, 1994; Muthén & Satorra, 1995; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999; Stapleton, 2002, 2006a, b; Yuan & Bentler, 2006). Most applied researchers are now aware that clustered data require special attention during analysis, such as the use of multilevel techniques. Therefore, non-independence of samples due to clustering will not be explored in this research.

One of the sampling design issues that has received less attention in the methodological research of Structural Equation Modeling (SEM) is stratification and unequal probabilities of selection. When sampling designs include stratification with unequal probabilities of selection, the goal is usually to ensure sufficient sample sizes for subgroup analysis.

There is documented evidence of the effects of ignoring unequal selection probabilities with statistical techniques such as regression (e.g., DuMouchel & Duncan, 1983; Nathan & Holt, 1980), ANOVA (e.g., Potthoff, Woodbury, & Manton, 1992; Wedel, Hofstede, & Steenkamp, 1977), single-level (e.g., Hahs, 2003; Kaplan & Ferguson, 1999; Stapleton, 2006a) and multilevel SEM (e.g., Lohr, 1999; Stapleton, 2002), growth models (Asparouhov, 2005), and latent class analysis (e.g., Patterson, Dayton, & Graubard, 2002). This research has determined that if unequal selection probabilities are not accounted for, resulting population estimates may exhibit substantial bias (Bentler & Chou, 1987; Lee et al., 1989; Skinner, Holt, & Smith, 1989). However, no research has yet documented the effects of sampling with unequal selection probabilities on GMM.

1.2 SIGNIFICANCE

Although Latent Growth Modeling has been used with considerable success in a wide variety of research settings, Growth Mixture Modeling is an important new development for applied researchers. A limitation of Latent Growth Models is their assumption that all individuals are drawn from a single observed population with common growth parameters (e.g., slopes and intercepts). However, in many applied situations, unobserved subpopulations may exist and their trajectories may be used to test a wide variety of theories and could also resolve inconsistent findings from prior research.

In fact, as researchers begin to use GMM with data from nationally representative samples, it becomes increasingly critical for researchers to understand the difficulties associated with modeling data that come from complex sample designs. Unfortunately, the sampling process is often ignored when it comes time for large scale data sets, and ignoring the dependencies in complex data may result in estimates that are not robust. This tendency to naively model complex data as if it were from a simple random sample (SRS) may be due to unfamiliarity with the effects of complex sample designs on subsequent parameter estimates and their associated standard errors. As long ago as 1974, Kish and Frankel stated that “in most cases the kind of education that aspiring statisticians are given at university does not prepare them for the complexity of the data that they are going to handle” (p. 29). Fifteen years later, Skinner et al., (1989) described the importance of such training: “analytic procedures which take into account the population structure and the sample selection mechanism can change the objectives of the analysis and can have a substantial impact on the subsequent interpretation of results" (p. 285).

If researchers naively apply GMM to nationally representative data sets without adjusting for the way in which the sample was selected, the resulting parameter estimates, standard errors and tests of significant may not be trustworthy. This research study will therefore provide applied researchers with practical guidance on appropriate methods for accounting for these violations of statistical assumptions.

1.3 ORGANIZATION OF CHAPTERS

The remainder of this document is organized as follows. Chapter 2 provides a review of the literature on unequal sampling probabilities and GMM. Chapter 3 describes the design of the proposed research. Chapter 4 provides the results, while Chapter 5 discusses the results and provides recommendations.

CHAPTER 2: REVIEW OF THE LITERATURE

Available research on the characteristics of samples that utilize unequal sampling probabilities, and the research documenting their effects on Growth Mixture Modeling are reviewed in this chapter. Section 2.1 reviews the literature on disproportionate sampling and traditional approaches used to analyze such data. Section 2.2 presents the relevant literature on Growth Mixture Models.

2.1 UNEQUAL SAMPLING PROBABILITIES

For the first 30 years of the twentieth century, the major issue in survey research was designing sampling schemes that would accurately represent the population. In 1934 Jerzy Neyman created such a design (Neyman, 1934); he established the use of probability samples, randomization, and introduced the concepts of stratification and unequal probabilities of selection. Because the selection probabilities of all subjects in the population are known, Neyman's ideas allow researchers to make statistical inferences from a sample to the population, and furthermore allow for the calculation of standard errors. These days, most large-scale surveys employ sampling designs based on Neyman's ideas. The result is a "complex sample design" if it has at least one of the following attributes: "unequal probabilities of selection for different units; stratification; ….. or clustering" (Mislevy, 1991, p. 178). Complex sample designs differ from survey to survey, but all generally contain unequal selection probabilities. Unequal probabilities of selection result when subjects in the population are sampled at different rates (Kish, 1995), which often (but not always) occur as a result of stratification.

2.1.1 Stratification

In broad terms, stratified sampling consists of dividing the target population into S non-overlapping groups, called strata ($s=1,2,\ldots,S$). Separate samples are then selected from within each stratum. Stratification thus separates the population into a number of observed groups from which independent samples are drawn (Kish, 1995). Strata are defined using supplementary information, such as regional, demographic, or socioeconomic variables, that are available for all members of the population.

Nationally representative samples usually employ stratification in the selection of subjects because it ensures that all important domains are included in the sample. Furthermore, stratification can be more efficient than simple random sampling because subjects with similar characteristics are collected together within a stratum, resulting in a small within-strata variance relative to between-strata variance (Lehtonen & Pahkinen, 1994). Stratification can produce more precise estimates of population characteristics if the stratification variable is correlated with the outcome variable (Heeringa & Liu, 1997; Kalton, 1983). This is because subjects within strata tend to be more homogenous than subjects in the population as a whole. This reduction in variance in individual stratum often lead to a reduced variance for the estimate of the total population.

2.1.1.1 Equal Probabilities of Selection

If sampling within each stratum is proportional to the size of the stratum, it is called "proportionate stratification" or a "self weighting sample" because the selection probability ($p_s = n_s/N_s$), is constant for each stratum s, and is also a constant for any population member. That is, $p_1=p_2=p_3=p_s=P$, which is the overall inclusion probability. The population mean can thus be estimated by the simple mean of the sample cases (subjects must be sorted into separate strata for computing the variance properly, but this sorting is not necessary for the computation of the mean). Here, the variances of the sample estimates obtained via proportionate stratification are always smaller than an SRS sample of the same size (Kish, 1995). The variance is decreased to the extent that there is homogeneity within strata and that stratum means diverge. Because proportionate stratification does not sample with unequal sampling probabilities, it will not be considered further in this study. Instead the study will focus on disproportionate stratified sampling.

2.1.1.2 Unequal Probabilities of Selection

Unequal probabilities of selection usually result from stratification when the goal is to ensure a sufficient sample size for subgroup reporting (Lohr, 1999). For example, with the Early Childhood Longitudinal Study (ECLS), there was special interest in reporting accurate estimates for Asian/Pacific Islander students.

Those students were therefore sampled at a rate three times higher than the rate for other students (U.S. Department of Education, 2003). When subjects in a stratum are sampled with higher probabilities than other strata, those subjects are said to be "over-sampled" because they have greater representation than would occur in SRS. When all subjects in a stratum are included in the sample they are "sampled with certainty." Disproportionate stratified sampling can also be used to lower the cost of sampling by increasing the selection probability in strata with lower costs.

Many government agencies sponsor large-scale studies that utilize disproportionate stratification and are longitudinal in nature, such as the Early Childhood Longitudinal Study (ECLS), the National Educational Longitudinal Study (NELS), the National Longitudinal Study of the High School Class of 1972 (NLS), and the High School and Beyond Longitudinal Study (HS&B). All of these studies employ some form of stratification in their sample design (U.S. Department of Education, 2003). For these studies, the portions of the sample design that result in unequal selection probabilities are presented in Table 1.

Table 1: Features of stratification designs that result in unequal probabilities of selection in 4 national studies

	ECLS	**NELS**	**NLS**	**HS&B**
Stratification	2 variables	3 variables	2 variables	2 variables
Oversampling	API* & private schools	API, Hispanic & private schools	Low income & minority	High % minority & Catholic schools
Sampling w. certainty	24 largest counties	Students within smallest schools	--	Students within smallest schools

**API=Asian and Pacific Islanders*

When analyzing the data from studies such as those presented in Table 1, it is important to remember a fundamental assumption under which nearly all statistical models operate: that each subject in the population has an equal probability of being selected into the sample. This assumption is very important because when a simple random sample is used, descriptive statistics reflect the population from which the sample was drawn. However, when some subjects are sampled at different rates than others, the resulting estimates reflect the sample and not the population. In fact, the estimates for the sample can be very different from the population (Bentler & Chou, 1987; Hahs, 2003; Pfeffermann, 1993). This difference between sample and population is commonly known as ‘selection bias.’

Researchers have demonstrated that if the outcome variable is related to the variables used to stratify the sample, it is likely that the parameter estimates will be biased unless the disproportionate stratification is accounted for in the latent growth model (Kaplan & Ferguson, 1999; Korn & Graubard, 1995a, b; Lee et al., 1989). Growth Mixture Modeling will be described in more detail in Section 2.2, but it is first useful to review the statistical techniques that have traditionally been used to analyze data from samples collected with unequal probabilities of selection.

2.1.2 Traditional Analysis of Data from Stratified Samples

Data from stratified samples have traditionally been used for descriptive purposes (such as summary estimates of the population), and estimation procedures that account for unequal selection probabilities are well established for these purposes (Kish, 1995). Inferential statistics, in contrast, attempt to explain the processes that underlie these descriptive estimates (Skinner et al., 1989). Estimation methods for inferential purposes that take stratification into account have not been fully explored, but two methods have been suggested in the literature. One is weighting observations (long used in descriptive statistics) and the other is modeling the sample design by adding covariates to the model.

An analysis that adjusts estimates but does not explicitly model the sampling design is referred to as a 'design-based' method (Kalton, 1983). Design-based methods are primarily concerned with accurately estimating population characteristics (Skinner et al., 1989). Accounting for the sample design in this way is frequently used in descriptive statistics, since the quality of the population estimates depends heavily on how randomization is applied in selecting the sample. Design-based methods make inferences to a known and finite population--the sampling distribution of repeated samples generated by the sampling design (Lee et al., 1989; Pfeffermann, 1993). Thus, the finite population is of primary interest, and the analysis aims at finding estimates that are design-unbiased in repeated sampling.

Model-based approaches, in contrast, condition on sampling variables (such as the variables used to stratify the sample) to adjust for the effects of the sample design (Skinner et al., 1989). Traditionally, model-based researchers draw inferences to populations that are more general than the fixed finite population that gave rise to the sample. Model-based inference thus assumes that a sample is a convenience set of observations from a conceptual super-population. In the context of sample surveys, Deming and Stephan (1941) considered a super-population to be a hypothetical infinite population from which the finite population is itself a sample. The population parameters under the specified model are of primary interest, and the sample selection scheme is considered secondary to the inference. Consequently, the role of the sample design is de-emphasized, and statistical estimation uses the prediction approach under the specified model (Lee et al., 1989). This perspective holds that it is not necessary to adjust for the sampling design because the results are not being generalized to the finite population (Pfeffermann, 1993).

If the population is infinite, every individual has a selection probability of zero, with a resulting weight of infinity (Asparouhov, 2005). It is important to remember that for large finite populations, the difference between a finite population parameter and the corresponding super-population parameter is small, and thus inference on the finite population parameter also constitutes inference on the super-population parameter (Lehtonen & Pahkinen, 1994). In the following sections, design- and model-based approaches are discussed in more detail.

2.1.2.1 Design-Based Approaches

To date, most statistical developments for complex samples have focused on univariate, design-based methods, with an emphasis on using sampling weights to estimate population parameters and standard errors. The use of weights in sample estimates traces back more than 50 years to the work of Horvitz and Thompson (1952), who found that unbiased estimates of population totals can be obtained by using sample weights, when sampling probabilities are known. When estimates of population quantities (such as means) are weighted, they will be unbiased when the sample size is large.

In applied situations, weights are the cumulative result of a comparison between the population structure and the sample structure, including nonresponse adjustments, stratification, post-stratification, and sampling design considerations. The generation of sampling weights can therefore be very complex. The goal of this dissertation is not to imitate such complex processes, but to demonstrate the fundamental principles underlying the analysis of data obtained with unequal selection probabilities. In fact, weights are usually adjusted upward to compensate for nonresponse and post-stratification to match known census figures, so the base weights used in the current research should be able to be generalized to more complex, final weights.

As described in Kish (1995), disproportionate sampling within strata necessitates the computation of weights for each subject, giving each stratum the same relative importance that it has in the population. Each sampled subject (i=1,2,...,n) is assigned a raw weight that is the inverse of the subject's selection probability, p_i, within their stratum, s. The selection probability is given as:

$$p_i = \frac{n_s}{N_s} \qquad (1)$$

where N_s is the population size of stratum s, and n_s is the corresponding sample size. Thus, if subject i has probability p_i of being included in a sample, then the raw weight w_i is given as:

$$w_i = \frac{1}{p_i} \quad . \tag{2}$$

The main purpose of weighting is to reduce bias in population estimates by up-weighting subgroups that are under-represented in the sample, and down-weighting subgroups that are over-represented in the sample. The raw weight for each subject can therefore be interpreted as the number of population members represented by that subject.

If y_i are scores on a measure with n subjects, under SRS an unbiased estimate of the population mean is defined as:

$$\mu = \frac{\sum_{=1}}{n} \quad . \tag{3}$$

And if the sample design includes unequal selection probability, the population mean is estimated by the weighted mean:

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i} \quad . \tag{4}$$

For a sample with equal probabilities of selection, the sampling variance of the estimator (Equation 3 or 4) is estimated by:

$$\sigma = \frac{\sum}{n-} \quad \overline{} \tag{5}$$

and the weighted sample element variance is calculated as:

$$\sigma_{..} = \frac{\sum}{\sum_{i=} \quad \sum_{=}} \quad \overline{} \quad . \tag{6}$$

Note that, while equation 4 is invariant with respect to constant multiplicative factors of weights, the weighted estimate of the sampling variance (equation 6) is not. It is therefore essential to scale the raw weights (Longford, 1995; Stapleton, 2002). Two methods of scaling exist, relative weighting and effective weighting. Relative weighting scales the raw weights so that they sum to the sample size:

$$w_r = w_i\left(\frac{n}{N}\right) \quad . \tag{7}$$

Effective weighting scales the weights so that their total is equal to the effective sample size (the sample size that would have been required to obtain the same level of precision as a simple random sample):

$$w_e = w_i\left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i^2}\right) \tag{8}$$

Here, the raw weights are scaled by multiplying them by the sum of the raw weights divided by the sum of the raw squared weights (Potthoff et al., 1992; Stapleton, 2002). For example, a population of 100,000 individuals is grouped into two strata (males and females) of 50,000 each, and each stratum is sampled at different rates. Eighty percent of the total sample of 500 will come from the first stratum and twenty percent will come from the second stratum. This sample design results in the following weights, as shown in Table 2.

Table 2: Example of relative and effective weights

Strata	**N_s**	**n_s**	**p_i**	**w_i**	**w_r**	**w_e**
1	50,000	400	.008	125	0.625	0.40
2	50,000	100	.002	500	2.500	1.60
Sum	*N=100,000*	*n=500*		*n=100,000*	*n=500*	*n=320*

2.1.2.2 Design-Based vs. Model-Based Approaches

Estimation of population means from data sampled with unequal probabilities of inclusion was studied by Little (1983), and in regression modeling by Holt, Smith, and Winter (1980), and Nathan and Holt (1980). For regression models, these researchers compared the results from a model-based design and results from a design-based approach to a baseline model that did not attempt to adjust for unequal probabilities of selection. They determined that without any adjustment, the resulting parameter estimates were biased. Both the model- and design-based estimates yielded more accurate estimates but the researchers preferred the model-based method because it tended to yield smaller variance estimates. However, if the model is incorrectly specified or the sampling design is too complex to model, this advantage is lost (Muthén & Satorra, 1995). It is important to note, furthermore, that both studies used raw weights in their design-based analysis, rather than scaled weights (i.e., relative or effective weights). The choice of approach thus comes down to "a tradeoff between the potentially larger bias of unweighted estimators and the potential larger variability of weighted estimates" (Korn & Graubard, 1995b, p. 295). Many researchers (e.g., Pfeffermann, 1993; Sturgis, 2004; Sugden & Smith, 1984) have determined that more accurate parameter estimates are preferred to an increase in precision.

Some advocates of the model-based approach criticize the use of design-based analysis because it ignores the fact that estimating a population total, *Y,* is equivalent to predicting the total of *Y* for the individuals who have not been selected into the sample. Hence, it is asserted, assumptions must be made that relate those individuals in the sample to those who are not in the sample, so that any inference about those not in the sample will be meaningful. If such relations exist, and are known, they assert that incorporating adjustments for the probability of selection is not necessary (Hansen, Madow, & Tepping, 1983). This criticism is related to another that asserts that when the sample is selected, all the researcher has is a unique sample, and that the selection process should therefore be ignored. In this view there is no model that provides the relationship between the sample and the balance of the population, and how that sample was selected cannot create the relationship. That is, inferences should not depend on the sample design, and that the design of the sample in such instances should be ignored.

The analysis is done as if the only source of variation were random sampling from a hypothetical super-population.

However, there are practical problems with including all of the necessary design variables into the model; often researchers do not have information on all of the relevant design variables for a particular sample (due to confidentiality concerns), or there may be too many variables to be incorporated into the model (Potthoff et al., 1992). “No model will include all of the relevant variables and few analysts will wish to include in the model all the geographic and operational variables which determine sampling rates. The theoretical and empirical tasks of deriving, fitting, and validating such models seem formidable for many complex national demographic surveys” (Pfeffermann, 1993 p. 326).

2.1.2.3 Approaches for Latent Variable Models

Despite a long tradition of use with multiple regression (e.g., DuMouchel & Duncan, 1983) and ANOVA (e.g., Muthén & Satorra, 1995; Potthoff et al., 1992), methods that incorporate adjustments for disproportionate sampling in latent variable models are limited. None could be found that deal specifically with Growth Mixture Models.

Muthén and Satorra (1995) were the first to address complex sample issues within a latent variable framework. They investigated both model- and design-based methods in a factor analytic model. They described the model-based approach as an efficient way to disentangle population heterogeneity. The main drawback to the approach, however, was that it provided accurate estimates only when the variables were normally distributed and the factor model was correctly specified: “While model-based disaggregated modeling has much to offer in terms of efficient estimation, it also suffers from a lack of robustness to misspecification….It is unlikely that exactly the same covariance structure model holds within different strata” (Muthén & Satorra, 1995, p. 295).

Kaplan and Ferguson (1999) explored using raw and relative sample weights on a single-level structural equation model. They found that when unequal probabilities of selection were ignored, it resulted in serious bias of the parameter estimates, with the most serious biases occurring in the factor variance and error variances. The use of raw and relative weights lessened the bias in these estimates, and bias diminished even more as strata sample size increased. Kaplan and Ferguson also found that raw and relative weighting systematically underestimated the standard errors relative to the standard deviation of the estimates. For the no weight condition, the biases tended to be positive, but uniformly smaller. The bias values for the raw weight conditions were much larger compared to the relative weighting condition because the raw weights sum to the population size.

Hahs (2003), in her use of sampling weights in a single level structural equation model, also found that ignoring weights resulted in parameter estimates that were significantly different from those obtained from the weighted sample, with the structural parameters evidencing the greatest bias. Unlike Kaplan and Ferguson (1999), however, she found that the weighted model standard errors were higher than the non-weighted estimates. However, Hahs used effective weights, whereas Kaplan and Ferguson used raw and relative weights. Furthermore, Hahs' investigation did not use a Monte Carlo methodology, but instead used existing data.

In multilevel modeling, which uses a model-based procedure to account for the effects of clustering, sample weights are sometimes used to account for unequal probabilities of selection. Stapleton (2002) found that when weights are not used in the calculation of the input covariance matrices, parameter point estimates were biased. However, even when raw weights are used, Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998), and Stapleton (2002) found that parameter estimates could still be biased. Stapleton (2002) determined that relative weights can result in "drastically underestimated sampling variances of parameters" (p. 497); up to 40% in some cases. Both Pfeffermann et al. (1998) and Stapleton (2002) found that greater bias occurred when the sampling weights were more unequal. When effective weights were used to calculate covariance matrices, however, they generated accurate parameter estimates and robust standard errors (Stapleton, 2002).

Patterson (1998) examined the importance of accounting for complex sample designs in latent class modeling (LCM). She found that when weights were not used, point estimates were underestimated by 10 to 30%. Patterson et al. (2002) also found that ignoring sample weights resulted in biased parameter estimates. They found that weighting can affect the estimation of the item-conditional probabilities and the latent class proportions when sampling weights differ across subgroups.

Furthermore, Asparouhov (2005) found that if unequal probabilities of selection were not adjusted for in a two class model, almost all parameter estimates contained substantial bias. “It is clear from our simulations that omitting the weights can produce severely biased estimates for any latent variable model. No parameter appears to be immune from selection bias” (p. 20).

In summary, when weights are not used to adjust for disproportionate sampling, the resulting parameter estimates are likely to be biased. Regardless of the philosophical approach (design- or model-based) taken, when a researcher attempts to connect the sample to the population, the literature suggests that it is necessary to account for how the sample is selected.

2.2 GROWTH MIXTURE MODELING

Some authors have described Growth Mixture Models (GMM) as a hybrid model that may be viewed as an extension of traditional continuous and discrete latent variable models (Bauer & Curran, 2004; Arminger & Stein, 1997; Muthén, 2002). GMM can thus be seen as simultaneously estimating two submodels, one with continuous latent variables (Latent Growth Model), and one with a categorical latent variable. GMM thus have their basis in both the common factor model of Thurstone (1947) and in the latent class model of Gibson (1959). It is useful to consider both of these ‘first generation’ submodels in turn before exploring GMM in more detail.

2.2.1 Submodel 1: Continuous Latent Variables

The correlation coefficient has long been interpreted as a potential indicator of the degree to which two variables share a common cause (e.g., Galton, 1888; Stigler, 1986). Factor analysis was developed as a tool for identifying these unobserved common causes (Spearman, 1904; Thurstone, 1935, 1947). The factor analytic model usually assumes that all of the shared variance among a set of variables are a result of the influence of one or more continuously distributed latent factors. After the effect of these common latent factors are accounted for, it is assumed that any residual relationship among the observed variables approaches zero (within sampling error). This is the assumption of conditional independence. Another assumption of factor analysis is that the observed variables can be expressed as a linear combination of the latent factors plus the residuals, and that the latent factors and the residuals are multivariate normally distributed.

2.2.1.1 Latent Growth Modeling

Growth models were initially developed by showing how factor analysis provided a framework for representing inter-individual differences over time (McArdle & Epstein, 1987; Meredith & Tisak, 1990; Muthén, 1991; Willett & Sayer, 1994). Latent Growth Models (LGM) are therefore a special type of confirmatory factor model, sometimes with a mean structure, where the latent factors represent the components of individual change over time (the intercept and slope). LGM allow the parameter values of these growth factors to vary for each individual, but assumes that the trajectories of all subjects follow the same functional form (e.g., linear, quadratic). Muthén and colleagues extended this basis to incorporate dichotomous variables and missing data (Muthén, 1992, 2001a, 2002; Muthén & Satorra, 1995).

The measurement portion of a latent growth curve model, with *n* subjects (*i*=1,2,…,*n*) measured on the continuous *y* (*y*=1,2,…,*p*) observed variables and with *m* latent growth factors (*m*=1,2,…*M*) can be expressed as:

$$y = \Lambda\eta + \varepsilon \tag{9}$$

where *y* is a $p \times 1$ vector of repeated measures and η is an $M \times 1$ vector of latent growth factors, and ε is a $p \times 1$ vector of residual errors. Λ is a $p \times M$ parameter matrix of factor loadings. By constraining the loadings in Λ to known constants, the latent intercept and slope factors are defined. The loadings from the intercept factor to each of the measured time points are set to 1. The paths from the latent slope factor to each of the time points are generally set to numbers representing an assumed linear or curvilinear trajectory. For example, for a linear assumption with four equally spaced time points, the slope loadings would be set to 0, 1, 2, 3. The measure of time is thus entered into the model via the factor loading matrix.

The latent portion of a latent growth curve model (without covariates) can be written as:

$$\eta = \kappa + \zeta \tag{10}$$

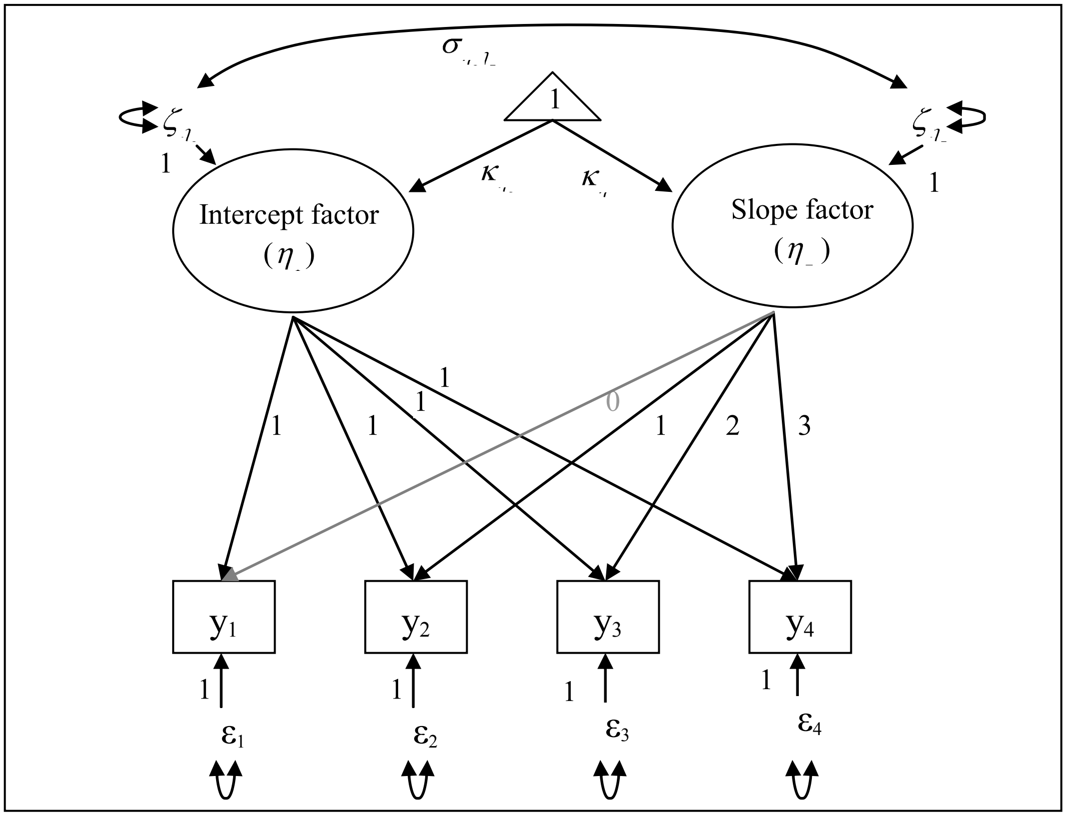
where κ is an $M \times 1$ vector of latent variable means and ζ is the $M \times 1$ vector of individual deviations from these means. Φ is the $M \times M$ covariance matrix of the latent factors, which allows the intercept and slope to be related to one another. When exogenous predictors (*x*=1,2,…*X*) are added, the model is rewritten:

$$\eta = \iota + \; x + \; \text{´} \qquad (11)$$

where γ is a $M \times X$ matrix of regression parameters between the *M* latent growth factors and the exogenous predictors. Given the regression of η on *x* via γ, α now contains the intercepts of the latent growth factors when *x*=0. ζ contains the individually varying residuals.

Conventional LGM may be estimated by maximum likelihood under normality assumptions using standard structural equation modeling software. A path diagram with equally spaced time points and linear growth is provided in Figure 1.

Figure 1: Conventional latent growth model with linear trajectory and mean factor



Unfortunately, LGM require an assumption that may restrict its applicability. LGM assume that all individuals in the sample come from a single population, with individual variation around a mean growth curve captured by the growth factor variances (Muthén & Muthén, 2000). Conventional LGM, therefore, cannot capture heterogeneity that corresponds to different subpopulations characterized by distinct growth trajectories. When a sample contains two or more qualitatively distinct, but latent, populations, important differences between the subpopulations are subsequently lost when their distributions are aggregated. This restriction is not present in GMM because they contain a second submodel which allows a categorical latent variable to be included in the model.

2.2.2 Submodel 2: Categorical Latent Variables

When the latent factor is categorical, Latent Profile Analysis (LPA) is used for continuous outcome variables, while Latent Class Analysis (LCA) is used when the outcome variables are categorical. The objective of LCA is to categorize individuals into latent classes and to identify items that best distinguish between classes. Both LPA and LCA are based on a different interpretation of the correlation coefficient than latent variable modeling with continuous latent variables. This alternative interpretation assumes that the correlation reflects the presence of discrete groups in the population, each characterized by either (a) different mean levels on the observed variables for LPA (Bauer & Curran, 2004), or (b) different probabilities of the categorical outcome variables for LCA (Lazarsfeld & Henry, 1968).

LPA was developed by Gibson (1959) as a way to identify these latent groups, and was presented as an alternative explanation to Thurstone's (1947) factor model. Although the interpretation of the correlation coefficient is different (with Thurstone's interpretation being that the correlation represents the degree to which two variables share a common cause, and Gibson stating that correlation indicates the presence of discrete groups in the populations), both factor analysis and LPA share the assumption of local independence. Gibson stated that, conditional on class membership, the residual association between the observed variables would be zero, within sampling variability. In LCA, local independence means that the mean of each latent class represents the true score of all individuals in that class. Within each class, any deviation from this mean should therefore be random and independent (Goodman, 1974; Lazarsfeld & Henry, 1968).

The total variance of the observed variables can be decomposed into between-class and within-class components used in conventional analysis of variance models, as shown in equation 12, below. The difference is that group membership is unobserved and must be estimated.

$$\sigma_y^2 = \sum_{k=1}^{K} \pi_k (\mu_{yk} - \mu_y)^2 + \sum_{k=1}^{K} \pi_k \sigma_{yk}^2 \qquad (12)$$

where y is an observed variable and k designates a specific latent class (k=1,2,3,…K), so μ_{yk} is the mean and σ^2_{yk} is the variance for variable y in class k. π_k indicates the proportion of cases belonging to class k. The covariance between two observed variables (y_1 and y_2) is:

$$\sigma_{y_1y_2} = \sum_{k=1}^{K} \pi_k (\mu_{y_1y_2} - \mu_{y_1})(\mu_{y_1y_2} - \mu_{y_2}) \qquad (13)$$

Note that there is no within class component to equation 13 (in contrast to equation 12). This means that any association between variables y_1 and y_2 must be accounted for by the between class component (the mean differences between the classes). Whereas the factor analytic model discussed previously decomposed the covariance matrix into continuous factors and residual variances, the latent class model decomposes the same covariance matrix into mean differences between latent classes and within class residual variances.

In fact, under certain conditions, the factor analytic model and the latent class model are equivalent. A factor analytic model with *M* factors is equivalent to a $K = M + 1$ class model (see Bartholomew, 1987; Gibson, 1959). Because of this, it has been suggested that factor analysis decomposes the covariance matrix to highlight the relationship among the variables, whereas models with categorical latent variables highlight the relationship among individuals (Bauer & Curran, 2004).

2.2.3 Growth Mixture Models

Although the conventional Latent Growth Model presented in section 2.2.1.1 allows for individual differences by estimating both a mean trajectory and the variation around this trajectory, its main drawback is that it assumes that all individuals come from a single population and share the same growth pattern. GMM relax the single population assumption to allow for differences across unobserved subgroups. This is accomplished by using a categorical latent variable to represent two or more distinct trajectory classes. The combined use of continuous and categorical latent variables allows individuals to vary around the mean growth curve for their particular subgroup (Bauer & Curran, 2004; Muthén, 2001a, b; Muthén & Shedden, 1999; Nagin, 1999).

Growth Mixture Models were introduced in the late 1990s by Muthén and Nagin (Muthén, Brown, Khoo, Yang, & Jo, 1998; Muthén & Shedden, 1999;

Nagin, 1999), and extensions to the framework were made by Muthén (2001a, 2003). Growth Mixture Modeling is a technique that is often used to test theories that subgroups with qualitatively distinct trajectories exist within a population. Examples include identifying different types of alcohol users (Muthén & Muthén, 2000), physical aggression (Nagin & Tremblay, 1999) and reading skills (Muthén, Khoo, Francis, & Boscardin, 2003).

In Growth Mixture Modeling, the data are assumed to come from *K* populations (*k*=1,2,…,*K*), with a categorical latent variable indicating the membership of each individual into one of the population groups, where *k* designates group and indicates that parameters may differ over groups. Conditional on class *k*, the measurement portion of the model can be expressed as:

$$y_k = \Lambda_k \eta_k + \varepsilon_k \qquad (14)$$

The specification of equation 14 is similar to that given for conventional LGM specified in equation 9. The equation for the growth parameter portion of the model across *K* classes is as follows:

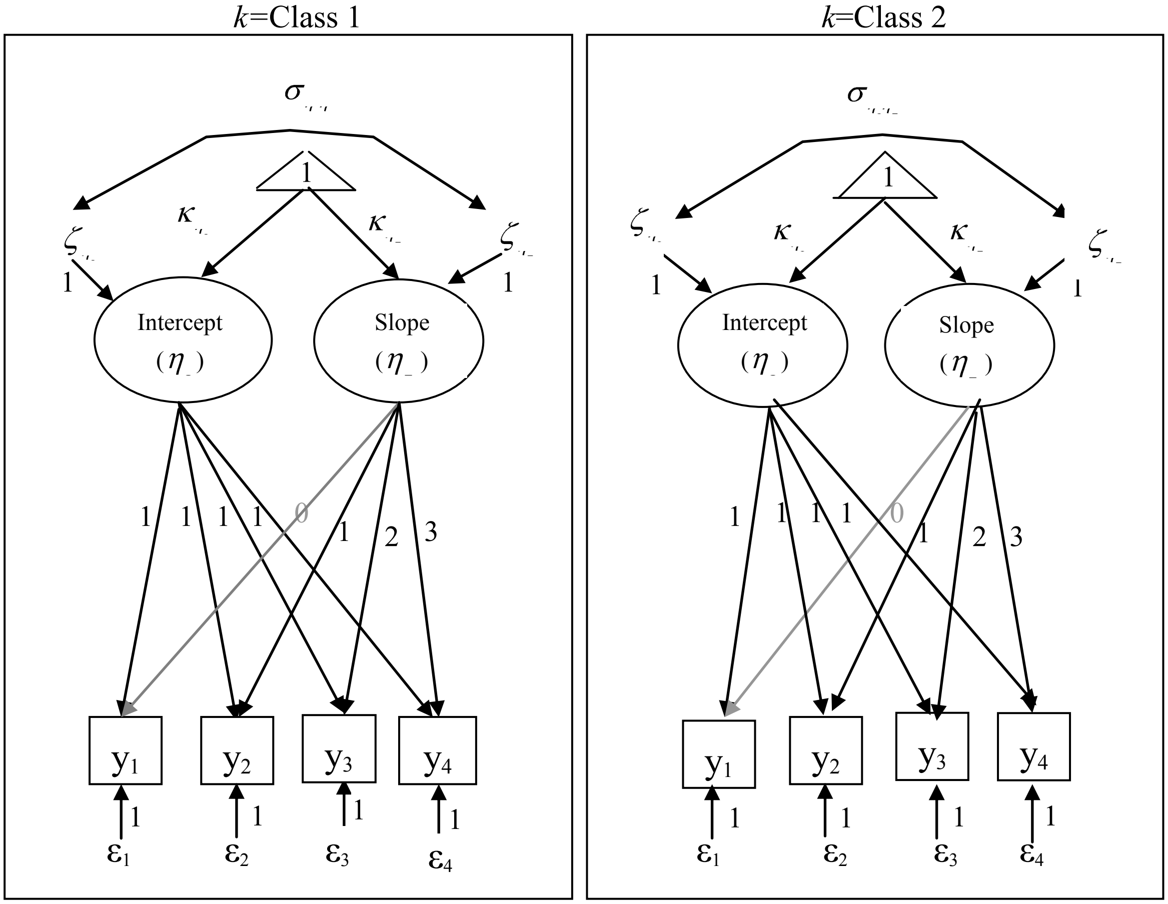
$$\eta_k = \alpha_k + \zeta_k \qquad (15)$$

The unconditional model in equations 14 and 15 simply extend the conventional Latent Growth Model expressed in equations 9 and 10 by having a separate equation model for each latent class, *k*, thus allowing for heterogeneity within the population. This Growth Mixture Modeling framework is also flexible enough to allow differences in measurement error variances across classes as well. Latent class membership may be related to other observed variables, so when covariates are added to the model, equation 15 becomes:

$$\eta = \chi + \gamma x + \zeta_{k} \tag{16}$$

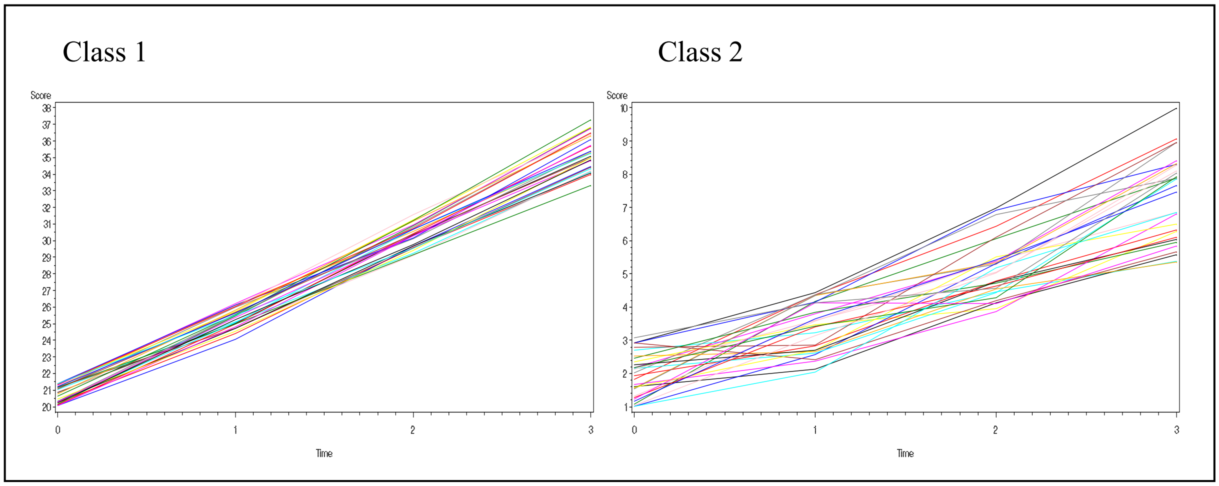
which, conditional on class membership, is equivalent to equation 11 of the conventional Latent Growth Model. An unconditional linear Growth Mixture Model with 4 equally spaced time points, a linear trajectory and two latent classes is displayed graphically in Figure 2.

Figure 2: Growth Mixture Model with 2 latent classes



Unlike the Latent Growth Model shown in Figure 1, the Growth Mixture Model in Figure 2 includes a categorical latent class factor where parameters associated with the growth factors vary across the different subgroups. The resulting trajectories from the 2-class Growth Mixture Model shown in Figure 2 could look similar to Figure 3 where class 1 has a relatively steep and compact trajectory with a high initial value, while class 2 has a flatter and more variable trajectory, with a lower value at the initial time point. Please note that in Figure 3, the scale of the y axes differ from class 1 to class 2.

Figure 3: Simulated trajectories of 2 latent classes



The Growth Mixture Model in Figure 2 contains the same 14 pieces of information (4 variances, 6 covariances and 4 means) as the conventional Latent Growth Model in Figure 1. However, the issue of model identification in Growth Mixture Modeling is more complex than in conventional Latent Growth Model (Li, Duncan, Duncan, & Acock, 2001) because a set of parameters are estimated for each class, resulting in many more estimated parameters than Latent Growth Modeling. In fact, Muthén and Shedden (1999) stated that “it is difficult to give rules for the identification of the (GMM) model.” However, as a well known example drawing on factor analysis demonstrates (Lawley & Maxwell, 1971), at least m^2 restrictions need to be imposed on the parameters and/or covariance matrix, when m is the number of latent continuous variables. As a result, GMM are generally not identified without implementing some parameter restrictions

(Muthèn & Shedden, 1999), such as constraining all of the error variances to be equal across classes.

Because class membership is unobserved, the proportion of individuals in each latent class is unknown and has to be estimated along with the other model parameters. Probabilities of group membership (posterior probabilities) are therefore estimated for each individual. These probabilities are used to weight each individual's contribution to the estimation of the model parameters for each class. Each individual's data thus contribute to the parameter estimates for each latent class commensurate with his/her probability of membership in that class (Muthén & Shedden, 1999).

While class probabilities are estimated, the number of latent classes must be specified by the researcher and are not estimated. Given that the groupings are latent, and therefore not known by the researcher, it is common to fit several models with different numbers of classes. For example, Boscardin's (2001) investigation into children's growth in pre-literacy skills fit six different Growth Mixture Models to the data, each containing a different number of latent classes, before determining that five distinct trajectory classes were present. To determine which model (and thus the correct number of classes) is appropriate; the models must be evaluated and compared in some way. Unfortunately, there is not one commonly accepted statistical indicator for deciding on the number of classes in a study population. In fact, a large body of research has focused on developing indices that may serve to compare alternate models (e.g., Bamber & Santer, 2000; Everitt & Hand, 1981; Lo et al., 2001; McLachlan & Peel, 2000; Nagin, 1999;

Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2008). These are discussed in following section.

2.2.3.1 Model Assessment

Model assessment is more complex in Growth Mixture Modeling than in conventional Latent Growth Modeling. For mixture modeling, the usual chi-square-based fit indices (e.g. CFI, RMSEA) cannot be used because there is not a single covariance matrix to which to fit the data—the saturated model cannot be estimated (McLachlan & Peel, 2000). Therefore, several other strategies have been proposed to evaluate model fit—namely: information criteria, number of classes, and the degree to which individuals are reliably classified into classes. However, there is not a commonly accepted criterion for determining model fit.

Information criteria (e.g., BIC) are frequently used for comparing models with different numbers of classes. Information criteria can be used as a basis for selecting the optimal model (and thus the correct number of classes). Usually, the smaller the information criteria, the better fit the model has to the data. However, Wedel et al. (1997) demonstrated that, if disproportionate sampling probabilities are not taken into account in the formulation of the likelihood, the selection of the appropriate number of classes will be affected.

As an alternative, Lo et al. (2001) proposed an empirical adjustment to the likelihood ratio statistic to test whether a particular sample is drawn from a mixture of *K* latent classes or from a mixture distribution of *K*-1 classes. This Lo-Mendall-Rubin (LMR) test compares the improvement in fit between neighboring class models (i.e., comparing *K* and *K*+1 class models) and provides a *p* value that can be used to determine if there is a significant improvement in fit when one more class is added. The more restrictive model is obtained from the less restrictive by a parameter assuming a value on the border of the admissible parameter space. Their technique derives the correct distribution as a weighted sum of independent chi-squared random variables with one degree of freedom (assuming within-class normality). Low *p* indicates that the *K*-1 class model should be rejected in favor of the *K* class model. This method has been described as a “breakthrough for helping to select the best-fitting number of classes” (Muthén, 2003, p. 7). However, an important criticism by Jeffries (2003) followed. Jeffries claimed that there is a flaw in the mathematical proof of the LMR test for nonnormal outcomes. He contended that this test requires conditions that are generally not met when the null hypothesis holds, and demonstrated that the log-likelihood ratios generated under the null hypothesis do not follow an approximately normal distribution.

Another way to gauge the fit of a Growth Mixture Model is assessing the ability of the model to assign individuals to their correct latent class. Entropy is a summary measure of classification accuracy that is provided as part of the Mplus output for mixture models (Muthén & Muthén, 2005). The formula for this entropy measure is:

$$E_K = -\sum\sum - \frac{\hat{p}_{ik}\ln\hat{p}_{ik})}{n\ln K} \tag{17}$$

where $\hat{p}_{ik}$ denotes the estimated conditional probability for individual i in class k. Entropy values range from 0 to 1 where values close to 1 indicate clear classifications (Muthén et al., 2002). In previous research, entropy values higher than .80 have been viewed as suggesting good classification (Greenbaum et al., 2005; Muthén, 2004; Wang & Bodner, 2007). In fact, the research of Lubke and Muthén (2007) indicates that entropy values below .60 are generally related to misclassifying approximately 20% or more of the sample. Entropy values around .80 and above are related to 90% correct assignment.

2.2.3.2 Distributional Assumptions

In Growth Mixture Modeling, latent classes are defined at the distribution level. Each population is composed of a collection of individuals who differ in their individual scores, but who are homogenous with respect to the population distribution from which they were sampled. Specifically, their observed scores are drawn from a common multivariate normal distribution, such that the same structural relationships hold for all individuals in the population. Population heterogeneity is therefore indicated by the presence of two or more latent groups in the population categorized by different distributions (Bauer & Curran, 2004).

Bauer and Curran (2003) demonstrated that a non-normal distribution is necessary for estimating the parameters of GMM. If the distribution of the sample is normal, then only one latent class is necessary, and there is no remaining information from which to identify a second class. That is, if the distribution of the repeated measure is multivariate normal, it should not be possible to obtain a multiclass solution when fitting a correctly specified growth model. With the correct model, the implied means and covariances of a single class should fully reproduce the observed distribution and additional classes should not be necessary. Non-normal sample distributions do not necessarily mean that multiple classes are present, however. The non-normality may be a symptom of either 1) the aggregation of distributions from several subpopulations, or 2) a non-normal distribution from one population. These two options are presented in Figures 4 and 5. Latent classes extracted via Growth Mixture Modeling may thus represent either the true latent subgroups in the population or may approximate a single homogenous, but non-normal, population distribution.

Figure 4: Non-normal distribution due to presence of two populations

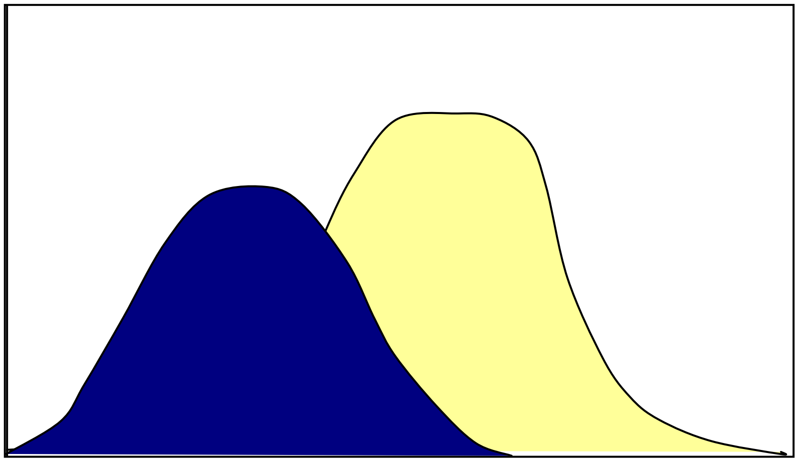
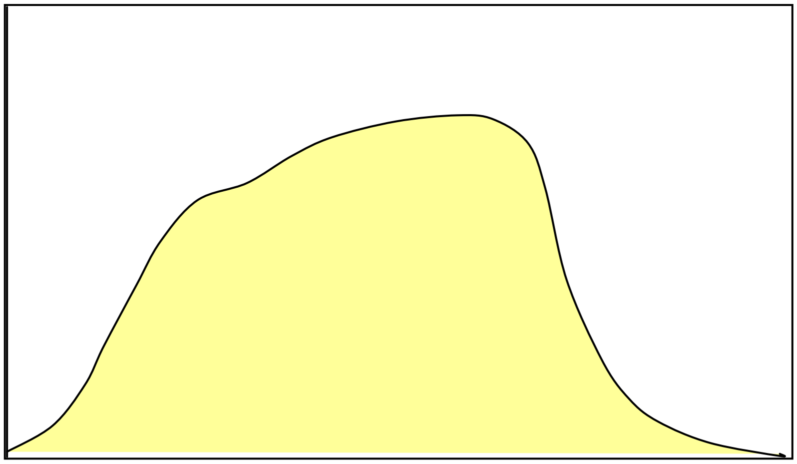


Figure 5: Nonnormal distribution of one population



Bauer and Curran in their simulation study (2004) found that the non-normality of the repeated measures was critical factor influencing both the estimation of a Growth Mixture Model and its fit relative to the correct (single) class model. Overall, their results supported their hypotheses that multiple latent trajectory classes would appear optimal for non-normality distributed data, even if these data were generated from a single homogenous population. In other words, a misspecified Growth Mixture Model would fit the data just as well as the correctly specified single class Latent Growth Model. However, the issue of equally well-fitting models is not new in Structural Equation Modeling—anytime a Structural Equation Model is fitted, it is well known that several other models will represent the sample data equally well. However, this problem of equivalent alternate models is confounded with the limitations of model assessment in Growth Mixture Modeling. While Bauer and Curran used BIC to determine model fit in their simulations, Muthén (2003) found that other fit indices would have better discriminated between a single population with a non-normal distribution and one that is due to the aggregation of multiple subpopulations.

2.2.4 Latent Classes and Stratification

At this point it is useful to compare the latent classes in Growth Mixture Modeling with the grouping of subjects created during stratification. As stated previously, latent classes in GMM are assumed to have been sampled from different (but unobserved) subpopulations. Stratification, in contrast, categorizes subjects by some observable trait or characteristic, such as region, employment status, or gender. Each subject in the sample thus belongs to two groups – the stratum from which they were sampled and their latent class, characterized by unobserved growth factors that describe their trajectory shape.

As an example, a researcher may be interested in how optimistic people are about the economy (and how this belief changes over time). The researcher will then identify an observable population of interest, such as all non-institutionalized adults in the state of Maryland. This researcher may then decide to segment this population into different strata based on an observed variable (such as employment status, educational attainment, or household income) before selecting her sample. Without knowing their latent subpopulation membership, the researcher independently selects a sample from each stratum. At this point, the researcher does not know how many unobserved subpopulations (latent classes) exist, and therefore does not know if class membership is related to the stratification variable. In this example, there could be a 2 latent classes present in the sample; individuals who are clinically depressed and those who are not.

If class membership is related to the stratification variable, one stratum would contain a higher proportion of a certain class than the others—for example, within the unemployed stratum there could be a significantly higher percentage of a depressed persons than in the corresponding employed stratum. Selecting subjects from one stratum at a higher rate would also therefore mean that subjects are also selected from one latent class at a higher rate – the extent depends on the degree of correlation.

When latent class membership is not related to the stratification variable, the subjects in each class would be randomly distributed among the strata. Therefore, with a large enough sample size, the expectation would be that each stratum would contain approximately equal proportions of each latent class. Unequal probabilities of selection that vary by strata should not affect the formation of latent classes.

2.3 RESEARCH GOALS

Most adjustments for data from complex samples have focused on univariate statistics and emphasize weighting the parameter estimates to correct for unequal sampling probabilities (Rust, 1985). In contrast, this study will instead focus on confirmatory latent growth models that include both continuous and categorical latent variables. The objective is to quantify the accuracy of parameter estimates and class assignment when subjects are sampled with unequal probabilities of selection. Specifically, this study:

1. Determines if, and under what circumstances, ignoring unequal sampling probabilities will result in biased parameter estimates and class assignments;
2. Develops and test methods for incorporating unequal sampling probabilities into GMM; and
3. Recommends the most practical and efficient procedure given the constraints of the field and available software.

To accomplish these goals, a series of Monte Carlo simulations empirically investigates the ability of GMM to recover known growth parameters of distinct populations when various adjustments are applied to the statistical model. Specifically, the study compares the performance of 1) GMM that ignore the sample design; 2) GMM that account for the sample design via weighting; 3) GMM that account for the sample design by explicitly modeling the stratification variable; and 4) GMM that account for the sample design by using weights and modeling the stratification variable. These models are compared across a variety of conditions which will allow the following research questions to be answered:

1. Can any of the GMM tested provide accurate parameter estimates when data are sampled with unequal selection probabilities from 2 latent populations?
2. Can any of the GMM tested reliably classify subjects into their latent class when data are sampled with unequal selection probabilities from 2 latent populations?

To answer these research questions, the following conditions are manipulated:

- Type of adjustment to GMM that accounts for unequal probability of selection: (4 conditions),
- Population characteristics (2 class proportion and 3 factor mean conditions), and
- Sample design (3 sample size and 3 probability of selection conditions).

These manipulated conditions (4×2×3×3×3) result in a Monte Carlo simulation with 216 cells. Detailed information on the research methodology, manipulated conditions, and descriptions of the latent variable models are provided in Chapter 3.

CHAPTER 3: STUDY METHODS

3.1 RESEARCH DESIGN

To investigate the performance of Growth Mixture Models when sample data are selected with unequal probabilities, a Monte Carlo simulation approach was used. Results from three Growth Mixture Models that use different methods to adjust for unequal probabilities of selection were compared to a standard Growth Mixture Model in which selection probabilities are ignored. To determine the extent to which the performance of these latent variable models is influenced by different population characteristics and sampling designs, these two conditions were manipulated. Parameter estimates were compared with their true values, and the percentage of properly converged replications and the precision of classification of class membership were calculated.

The remainder of this chapter describes the research design, including details on the fixed and manipulated conditions and the performance criteria used to evaluate the experimental models.

3.2 SIMULATION DESIGN

Conducting the simulations was a four-phase process. Phase 1 generated the mixed sample in SAS 9.1.3 (SAS Institute, 2004) using its Interactive Matrix Language (IML). This simulation assumed that the latent variable models describe a super-population, and that the finite population behaves like a sample from this infinite population. The model parameters thus define the super-population and are the target of inference (Nathan & Holt, 1980). As described in section 2.1.2, a super-population is a hypothetical infinite population from which the finite population is itself a sample. According to Skinner et al. (1989), “super-population parameters may often be preferred to finite population parameters as targets of inference in analytic surveys. However, if *n* is large, there will often be little numerical difference between the two.” (p. 14). Hence, in the remainder of this document, the distinction between finite and super-population parameters will not be emphasized.

For each cell in the study design, different seed values were specified to generate data. Seeds were used without replacement and were selected to be an odd nonnegative integer with a value less than 2^{31}-1 (or 2,147,483,647) (SAS Institute, 2004). Two multivariate normally distributed populations were generated in accordance with the manipulated conditions for each cell. These data were then combined into a single sample and exported out of SAS into Mplus 3.12 (Muthén & Muthén, 2005).

Phase 2 was the mixture analysis. The latent variable models were correctly specified and parameters estimated using Mplus. The default estimator for mixture analysis using Mplus is maximum likelihood with robust standard errors (Muthén & Muthén, 2005). Parameter values for the models are provided in Table 6. As this research was not intended to focus on convergence issues, true population parameters were used as the starting values in Mplus (Paxton, Curran, Bollen, Kirby, & Chen, 2001). Muthén and Muthén (2000) suggested that researchers provide starting values reflecting their beliefs about the population, as non-differentiated starting values may cause the modeling algorithm to fail. Furthermore, to investigate the differences between using the population values as starting values and using the Mplus default, five simulations were run both ways. These simulations indicated that there was no difference in parameter estimates between these two approaches.

In applied settings the population values are not known, and the default starting values in Mplus are often used (10 random sets of start values with two of the solutions with the highest log-likelihood from the EM algorithm chosen to be iterated until convergence). However, another useful way to minimize convergence problems is to use previous research to estimate appropriate starting values and to build up a model by estimating the model parts separately to obtain appropriate starting values for the full model.

For each of the experimental conditions, 1,000 properly converged replications (a replication for which the solution converges with no parameter estimates outside the possible range for the parameter) were generated. There was an upper limit of 20,000 on the number of attempted replications. In addition, a stop criterion was used such that after every 2,000 replications, if the percentage of properly converged solutions is less than one percent, the simulation ended for that cell. Cells that failed to achieve 1,000 properly converged replications were not considered for additional investigation (parameter estimates were not computed for these replications). Convergence for the remaining cells was measured by the number of replications needed to acquire 1,000 properly converged replications. A properly converged replication is one that converged to a solution according to the Mplus' default convergence criterion and had parameter estimates that were within the range of possible values (no negative variances). The default converge criterion for the Quasi-Newton algorithm in Mplus is .000001 (Muthèn & Muthèn, 2005).

In phase 3, SAS was again used, this time to import the parameters of interest from the Mplus output. In the fourth phase, SAS was used to compute the averages and variances across the successful replications in each cell and then this information was exported into excel files for table generation. In addition to a descriptive analysis, a decomposition of variance procedure was applied to the data to determine the proportion of variability explained by the manipulated conditions. To focus on the strongest relations, an effect had to explain at least 10% of the variability to be reported.

Some conditions were fixed throughout all simulations, while other conditions were manipulated. The fixed and manipulated conditions are described in the following sections.

3.2.1 Fixed Conditions

Conditions that were fixed across all simulations can be categorized into those pertaining to the latent variable model and those pertaining to population characteristics. These fixed conditions are described in more detail in the following sections.

3.2.1.1 Latent Variable Models

It has been demonstrated that the complexity of the model with respect to the factor structure, or the number of observed variables within class, do not influence model performance (Lubke & Muthén, 2007). However, mixture models in general are prone to local maxima of the likelihood (Bauer & Curran, 2003; Lubke, Muthén, & Larsen, 2002; Muthén, 2001a), which makes model convergence a potential issue for this study. The degree to which convergence is likely to be a problem is related to the information about the latent classes available in the data, and the complexity of the model. As this research is not intended to focus on convergence issues, the probability of encountering non-converging replications was minimized to the extent possible.

Because of the problems with model convergence, experts have recommended that that researchers build GMM iteratively, starting with the simplest model and adding parameters stepwise (Muthén, 2001a). Each of the five latent variable models tested will be correctly specified, single-level models, with linear growth trajectories. The models proposed for this study are relatively straightforward in this regard, and can be thought of as examples of the starting point of this iterative model-building procedure.

Furthermore, population parameters were used as the starting values in Mplus. The variances of the observed variable error terms were constrained to be equal across classes, the intercept and slope factors were uncorrelated, and their variances were constrained to be equal across classes. Constrained parameter values are provided in Table 3.

Table 3: Constrained parameter values

Growth Factor Variances	
Intercept	5
Slope	1
Growth Factor Covariance	
Intercept and slope	0
Manifest Variables	
Error variances	5,6,9,14

These constraints were implemented because, according to research, mixture models with large differences in the factor variances and covariances between classes are particularly sensitive to local maxima (Muthén, 2001a). Furthermore, in their simulations, Bauer and Curran (2003) found that allowing factor variances to vary did not predict class membership any more accurately than when factor variances were constrained across classes. Constraining factor variances to be equal across classes is furthermore common in the applied literature, such as in the study by DeFraine, VanDamme and Onghena (2004).

Observations at four equally spaced time points were generated. Applied studies that use four repeated observations include Kreisman (2003), Li, Barerra, Hops and Fisher (2002), and Muthén (2001b). Furthermore, according to research conducted by Lubke and Muthén (2007), additional time points do not influence model performance or class assignment.

As previously stated, the models under investigation were designed to be first-stage models in the iterative model-building process. Therefore an assumption of linear growth was implemented, with the assumption that additional polynomials would be added to the model at a later stage (if necessary). In the social sciences, the growth trajectory of first choice is frequently a simple linear function of time (Willett, 1988).

In many settings, theory has guided the determination of trajectory shape, with researchers choosing to model a linear trajectory. These studies include investigations of adolescent alcohol use (Duncan, Duncan, Strycker, Okut, & Li, 2002), and those investigating academic achievement (Kreisman, 2003). In fact, the mechanisms driving change over time are often unknown and therefore an empirical model-selection strategy usually leads to the adoption of a lower order polynomial to represent growth. Second, as pointed out by Willett and Sayer (1994) and Willett (1988), only a restricted portion of the growth is observed in most research, with a minimal number of waves of data collected. Therefore, only a growth function with a small number of parameters can be fitted. Moreover, even though growth may be curvilinear in the long run, it may be locally linear, which is another reason for a linear trajectory to be modeled.

3.2.1.2 Population Characteristics

The population always contained two latent classes. In the applied literature it is often determined that a 2-class solution provides the best fit to the data (e.g., DeFraine et al., 2004; Kreisman, 2003; Li et al., 2002; Muthén, 2001b). For simulation purposes, using only two classes kept the scope of the study manageable, but results should generalize to larger numbers of latent classes.

3.2.2 Manipulated Conditions

Manipulated conditions can be classified into three groups; those pertaining to 1) the latent variable model, 2) the characteristics of the populations from which the samples are drawn, and 3) the design used to sample the observations. The manipulated conditions are summarized in Table 4 before being described more fully in the following sections.

Table 4: Manipulated Conditions

Category	Number	Conditions
Latent variable models	4	- Baseline Model - Experimental Model I - Experimental Model II - Experimental Model III
Population characteristics	2	*Class proportions* - 50-50 - 70-30
	3	*Mean separation* - Intercept and slope - Intercept only - Slope only
Sample design	3	*Sample size* - 5,000 - 10,000 - 15,000
	3	*Sample selection* - SRS (no stratification) - Small differences in probability - Large differences in probability

The combination of manipulated factors results in 216 cells (4×2×3×3×3). These conditions are discussed in more detail in the following sections.

3.2.2.1 Latent Variable Models

Four latent variable models were tested in this investigation. They are:

1. Baseline Model, which ignores sample design;
2. Experimental Model I (Design-based), which uses an effective weighting strategy;
3. Experimental Model II (Model-based), where the stratification variable is modeled;
4. Experimental Model III, which combines Models I and II.

The Baseline Model assumes a simple random sample and does not incorporate strategies to account for unequal selection probabilities. This Baseline Model is widely used in practice, and corresponds to the default option in standard statistical packages.

In contrast, experimental Models I, II, and III attempt to compensate for the sample design in different ways. In Model I, weights were applied to the dataset (see Section 3.2.2.3 for more information on the weights used). Model II used the stratification variable as a covariate in the analysis, and weighting was not used. Model III incorporated both weights (as in Model I) and covariates (as in Model II). Model I thus corresponds to a design-based analysis; Model II to a model-based analysis; and Model III can be considered a combination of the design- and model-based approaches. In fact Skinner et al. (1989) stated that it is "sensible in practical applications to combine ideas and methods (of model- and design-based approaches)" (p. 10).

There is an ongoing debate in the literature as to whether it is appropriate for model-based methods to use weights in addition to covariates in their analysis (Longford, 1995; Nylund et al., 2007; Potthoff et al., 1992). According to Sugden and Smith (1984), if design variables have been incorporated into the model to reflect the stratification, and the model is correctly specified, then it is not necessary to weight. However, it is important to remember that, in applied situations, researchers don't know whether their model is correctly specified, and so weighting could be used as a hedge against model misspecification.

3.2.2.2 Population Characteristics

The proportion of individuals in each latent class, and the difference in the growth factor means between the two classes were manipulated in this experiment. First, two class populations were generated. The first condition set the proportion of subjects in both classes to be the same, so that the mixture in the aggregated population was 50:50. The other condition set one population to be larger than the other, with a 70:30 ratio in the aggregate population. This mixing proportion of 70:30 is seen in the applied literature (Zheng & Frey, 2004).

Mean separation (the standardized difference in latent means between the two classes) was also manipulated. These three manipulated conditions are provided in Table 5.

Table 5: Mean separation – standardized differences in latent means

		Standardized mean difference	
Condition	**Means differ**	**Intercept**	**Slope**
1	Intercept & slope	2	2
2	Intercept only	2	0
3	Slope only	0	2

The mean separation of factors from the first class to the second is 2 standard deviations, a value that is thought to provide well separated classes

(Nylund et al., 2007; Lubke & Muthén, 2007), and is similar to what is found in the applied literature (Greenbaum et al., 2005; van Lier et al., 2004). It is somewhat intuitive that when the means of the growth factors differ between classes, class assignment will be improved. However it is not known whether both growth factors have to differ, or if a mean difference in just one of the factors will suffice.

3.2.2.3 Sample Design

Two facets of the sample design were manipulated, the size of the sample and how the sample was selected from the population. Each of these condition types are discussed in more detail, below.

Simulation studies have shown that parameter recovery in structural equation models is affected by sample size (e.g., Gagné & Hancock, 2006), and it is reasonable to expect that Growth Mixture Models would be similarly affected. Therefore, three sample sizes were used to investigate their effect on the extraction of the true population values. The three sample sizes tested were 5,000, 10,000 and 15,000. This is larger than the sample sizes found in the applied literature; in 41 applied studies using mixture models the average sample was 909 subjects. However, the goal of the current research is to allow researchers to use GMM with nationally representative data sets, whose sample sizes tend to be considerably larger.

In terms of the sample selection procedures tested, the first sample drawn was a simple random sample to provide a baseline for the other two selection procedures. The remaining two sample conditions were drawn from a population that had been grouped into two strata. All individuals were categorized by a dichotomous observed variable (such as gender). This observed variable was used to stratify the population for sample selection, with the rates of selection from each of the two strata differing.

Brogan (1998) and Stapleton (2002) both found that greater bias in parameter estimates occurred in models where the sampling weights were more unequal. Therefore, the second condition employed a large difference in selection probability: with 80% of subjects from the first stratum selected, and 20% of subjects in the second stratum selected (4:1 ratio). The third condition used a small difference in selection probability, with 60% from the first stratum and 40% from the second stratum (1.5:1 ratio).

For each of the sample selection conditions, the probability of selection (p_i) is provided in Table 6, below.

Table 6: Manipulated conditions – Sample selection

Selection Probability Condition	***n****	p_i
SRS condition		
Total (no strata)	5,000	.10
80-20 condition (large difference)		
Stratum 1 (80%)	4,000	.08
Stratum 2 (20%)	1,000	.02
60-40 condition (small difference)		
Stratum 1 (60%)	3,000	.06
Stratum 2 (40%)	2,000	.04

** for sample size condition n=5,000*

3.2.2.4 Simulation Parameter Values

As stated previously, the factor variances and covariances were constrained to be equal across classes. In the population, each growth factor variance and their covariance had the same value in each class. In the Mplus code, these class 2 variance and covariance estimates were constrained to be equivalent to the class 1 estimates, and the single value was estimated. The intercept factor variance was set to 5, while the slope factor variance was set to 1. This variance ratio of 5 to 1 is commonly found in the literature (Muthèn & Muthèn, 2002). The covariance between the intercept and slope was set to 0, which is also a common condition in the applied literature (e.g., DeFraine et al., 2004; Kaplan, 2002).

The total variance of each of the four repeated measures was partitioned equally between the underlying growth factors and the time specific residuals. This methodology for determining the error variances was used by Bauer and Curran (2002). All parameter values specified in the simulation are provided in Table 7, below.

Table 7: Manipulated and fixed parameter values

		CLASS 1	**CLASS 2**
Fixed Conditions	**Growth Factor Variances**		
	Intercept	5	5
	Slope	1	1
	Growth Factor Covariance		
	Intercept and slope	0	0
	Manifest Variables		
	Error variances	5,6,9,14	5,6,9,14
Manipulated conditions	**Class separation**		
	I & S both differ		
	Intercept	2.24	6.72
	Slope	1	3
	Only I differs		
	Intercept	2.24	6.72
	Slope	1	1
	Only S differs		
	Intercept	2.24	2.24
	Slope	1	3

Many applied research studies that utilize GMM start with a constrained model where error variances and factor variances are constrained across classes. As the model building process progresses, these constraints are relaxed in an iterative process. As described previously, the models used in this study reflect the initial models used by researchers in the field (e.g., van Lier, Muthèn, van der Sar, & Crijnen, 2004; Weisner & Capaldi, 2003).

3.3 PERFORMANCE CRITERIA

Before performance criteria were calculated, data were checked to see if label switching occurred. Over the course of multiple iterations, the algorithm used by Mplus does not always assign 'Class 1' to the same group in the sample. If ignored, the estimated parameters for a given class will not to be meaningful. Although there is no established way to prevent label switching in Mplus, it is possible to tell if label switching has occurred. If the data obtained from the multiple iterations are sorted in descending order by a one of the estimated parameters, the mean will jump at a certain point and then continue roughly at the new level for the rest of the iterations. With label switching thus identified, these iterations were manually switched back to the appropriate class grouping.

In order to evaluate the performance of the models being tested, the following criteria were used: convergence rate, parameter bias, standard errors, and correct class assignment. The parameter estimates and class assignment were compared to known population values. However, population standard errors cannot be computed, so a comparison of the relative size of the estimated standard errors across conditions (within a given sample size) was made.

The convergence rate was examined to see if it appeared to be a function of the manipulated conditions. As stated previously, convergence was measured by the number of replications needed to acquire 1,000 properly converged replications. There was an upper limit of 20,000 on the number of attempted replications, as exceeding that limit would indicate that less than 5% of the replications were converging properly. Furthermore, a stop criterion was used such that after every 2,000 replications, if the percentage of properly converged solutions was less than 1%, then the simulation ended for that cell. Cells that failed to achieve 1,000 properly converged replications were not considered for additional investigation. Convergence for the rest of the cells was measured by the number of replications needed to acquire 1,000 properly converged replications.

A properly converged replication is one that converged to a solution according to the Mplus' default convergence criterion and had parameter estimates that were within the range of possible values (no negative variances). The default converge criterion for the Quasi-Newton algorithm in Mplus is .000001 (Muthèn & Muthèn, 2005).

As this research was not intended to focus on convergence issues, the true population parameters were used as the starting values in Mplus (Paxton, Curran, Bollen, Kirby, & Chen, 2001).

The accuracy of the parameter estimates was assessed by computing their proportional bias. The proportion of bias in the following parameter estimates was calculated; growth factor means, growth factor variances, and the mixing proportion. To do this, the population parameter value was subtracted from the average of the parameter estimates. This number was then be divided by the population value and multiplied by 100.

$$B(\hat{\theta}_p) = \frac{\bar{\theta}_p - \theta_p}{\theta_p} \times 00 \tag{18}$$

where θ_p is the population value for the *p*th parameter $\theta_p \neq)$, and $\bar{\hat{\theta}}_p$ is the mean of the estimates for the *p*th parameter across replications (Hoogland & Boomsma, 1998). Positive values for proportional bias indicate that the estimate is above the true value by the percent magnitude listed, whereas negative values indicate that the average estimate is below the true value by the percent magnitude. Hoogland and Boomsma (1998) indicated that this bias measure can be used to provide an indication of whether estimates are acceptable. They proposed that, for parameter estimates, bias measures within 5% of the population value could be considered acceptable.

Lastly, the extent to which subjects are accurately classified was assessed using the entropy measure described in Section 2.2.3. Entropy values range from 0 to 1 where values close to 1 indicate clear classifications; but values equal to or greater than .8 are considered acceptable (Muthén et al., 2002).

CHAPTER 4: RESULTS

Each of the 4 different approaches to adjusting for sample selection (The Baseline Model and the three Experimental Models) were simulated and the results are provided in this chapter. First, the Baseline Model results are described in terms of convergence, proportional bias, standard errors, and entropy. Next, the three experimental models (Models I, II and III) are compared to the Baseline Model in turn. These comparisons provide information on the performance of the three different methods being tested as adjustments for complex sampling designs (weighting, use of covariates, and both weighting and the use of covariates), compared to no adjustments at all.

Because true models were fitted to the data, the results provided in the following sections represent an upper bound of what be expected in an empirical study.

4.1 Baseline Model

The Baseline Model represents the status quo of Growth Mixture Modeling; it does not account for complex sampling designs. This section will first examine the results when simple random sampling (SRS) was used to generate the data. In this situation, the Baseline Model is used appropriately, as no adjustments for sample design are necessary. Next, the Baseline Model is applied to data that are generated from a complex sample. Here, the Baseline Model may be inappropriate, as stratification results in unequal probabilities of selection.

These two sets of results will be compared to estimate the impact of ignoring a complex sample design.

4.1.1 Model Convergence

When the sample selection process employed SRS, the Baseline Model attained a 100% convergence rate. However, when only one factor mean differed between classes, additional replications were sometimes needed to attain 1,000 converged replications. The number of additional replications needed is provided in Table 8. (All converged solutions were admissible with no negative estimates of variance.)

Table 8: Baseline Model - Number of additional replications needed when sample is SRS

Sample selection	Mean separation	Class proportion	Sample size	Additional iterations needed
SRS	Slope only	50-50	5,000	239
SRS	Slope only	70-30	5,000	191
SRS	Slope only	50-50	10,000	68
SRS	Slope only	70-30	10,000	63
SRS	Slope only	50-50	15,000	26
SRS	Slope only	70-30	15,000	18
SRS	Intercept only	50-50	5,000	11
SRS	Intercept only	70-30	5,000	3

When the sample procedure used was more complex, model convergence for the Baseline Model was also 100%. Moreover, when only one factor mean differed between classes, additional replications were again necessary. When the difference in the probability of selection was small (the 60-40 condition), four cells required additional replications. When the difference in the probability of selection was larger (the 80-20 condition), only two cells required additional replications, as shown in Table 9.

Table 9: Baseline Model - Number of additional replications needed when data are sampled disproportionally

Sample selection	**Mean separation**	**Class proportion**	**Sample size**	**Additional iterations needed**
60-40	Slope only	50-50	5,000	13
60-40	Slope only	70-30	5,000	44
60-40	Slope only	70-30	10,000	5
60-40	Slope only	70-30	15,000	1
80-20	Intercept only	50-50	5,000	14
80-20	Intercept only	70-30	5,000	3

All converged solutions were admissible with no negative estimates of variance.

4.1.2 Parameter Bias

The SRS condition for the Baseline Model illustrates a very simple scenario; the sample was chosen by simple random sampling, and an appropriate 2 class Growth Mixture Model was fitted. Table 10 provides the proportion of parameter bias for this initial scenario.

The first three columns of the table provide the experimental conditions. Column 1 gives the class size; whether the class sizes were equal (where each class contains 50% of cases), or unequal (where the first class contains 70% of cases and the second class contains the remaining 30%). The second column provides the mean separation condition. Here, ‘is’ indicates that both the intercept and slope factors in class 2 are different from those in class 1. Similarly, ‘i’ indicates that only the intercept factor mean differs, and ‘s’ indicates that only the slop mean differs. The third column provides the sample size conditions. The next 7 columns present the proportional bias found in the estimated parameters for each of the cells, given the experimental conditions. Lastly, the average proportion of bias across the estimated parameters is given, followed by the absolute value of the proportional bias. The difference between the average and the absolute value can be used as an indication of the variability in bias. Because the average bias uses negative values in its computation, the negative bias can cancel out the positive bias. Therefore, if the difference between the average and the absolute value average is small, most estimates of proportional bias were of the same sign. If the difference is large, however, the proportional bias estimates bounced around on both sides of zero.

Table 10: Baseline Model – Proportional bias by cell when sampling is SRS

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-0.089	0.066	0.149	0.242	0.050	0.064	-0.092	0.056	0.107	-0.052
equal	is	10,000	-0.011	0.112	0.051	0.190	0.011	0.023	-0.056	0.046	0.065	-0.019
equal	is	15,000	0.080	0.046	0.052	0.341	0.028	-0.007	0.045	0.083	0.085	-0.002
unequal	is	5,000	0.050	-0.093	3.853	4.739	0.010	0.033	-0.010	1.226	1.255	-0.029
unequal	is	10,000	0.030	-0.147	3.396	4.353	-0.002	0.025	-0.060	1.085	1.145	-0.060
unequal	is	15,000	-0.007	-0.060	2.877	3.197	-0.004	-0.036	-0.002	0.852	0.883	-0.031
equal	i	5,000	1.937	0.360	5.965	1.153	7.360	-0.259	-0.234	2.326	2.467	-0.141
equal	i	10,000	1.147	0.104	1.250	0.317	1.693	-0.092	-0.043	0.625	0.664	-0.039
equal	i	15,000	-0.019	0.052	-0.194	0.255	-0.028	-0.288	-0.118	-0.049	0.136	-0.185
unequal	i	5,000	2.637	0.536	11.016	0.855	8.308	0.879	-0.086	3.449	3.474	-0.024
unequal	i	10,000	3.223	0.186	6.030	-0.083	-2.001	1.663	0.017	1.291	1.886	-0.595
unequal	i	15,000	1.166	0.042	1.475	0.608	-1.302	0.408	-0.016	0.340	0.717	-0.377
equal	s	5,000	0.432	7.110	1.035	4.490	2.587	-0.093	0.021	2.226	2.253	-0.027
equal	s	10,000	0.052	3.581	0.221	1.258	0.330	0.258	-1.055	0.663	0.965	-0.301
equal	s	15,000	0.159	3.241	0.246	1.458	0.642	-0.027	-0.512	0.744	0.898	-0.154
unequal	s	5,000	0.985	9.012	1.394	6.615	-4.710	-0.227	5.822	2.699	4.109	-1.410
unequal	s	10,000	0.250	5.936	0.792	5.659	-5.163	-0.262	1.354	1.224	2.774	-1.550
unequal	s	15,000	0.322	4.899	0.375	4.644	-4.254	-0.360	1.261	0.984	2.302	-1.318

Table 10 shows that, with SRS, the percentage of bias in the parameter estimates is modest. Hoogland and Boomsma (1998) proposed that parameter estimates within 5% of the population value could be considered acceptable. Using this criterion, only 12 cells (9.5%) fell into the unacceptable range. The variance of the intercept and slope factors had the most bias, followed by the mixing parameter.

Looking at Table 10, the mean separation condition seems to affect the accuracy of parameter estimates. The parameters were estimated most accurately when both the intercept and slope factor means differed between classes.

The addition of a complex sampling structure (adding strata with different probabilities of selection) had an adverse affect on the accuracy of the Baseline Model. When the probability of selection varied only slightly between strata (individuals in strata one were sampled with a .06 probability, and strata two was sampled with a .04 probability: the 60-40 condition) the proportional bias was higher than when the sample was selected completely at random. While 9.5% of cells from the SRS condition fell into the unacceptable range (Hoogland & Boomsma, 1998), 50% of cells (63) were unacceptable when the sample selection probabilities varied in the 60-40 condition. When the probability of selection was manipulated to be more unequal (the 80-20 condition), the proportional bias increased, with 72% of cells (91) falling in the unacceptable range.

Tables 11 and 12 provide the proportional bias found in the parameters when the sample was stratified disproportionally. Table 11 shows the proportional bias when stratification resulted in a small difference in probability. Table 12 shows the same results for the large difference in probability condition.

Table 11: Baseline Model – Proportional bias when sample is stratified with small differences in the probability of selection

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-81.115	-20.819	-124.035	-121.402	-56.610	11.008	-64.881	-65.408	68.553	-133.961
equal	is	10,000	-80.719	-19.807	-141.228	-90.719	-56.026	10.767	-62.281	-62.859	65.935	-128.794
equal	is	15,000	-80.785	-19.720	-145.124	-83.782	-56.021	10.766	-62.024	-62.384	65.460	-127.845
unequal	is	5,000	-1.415	12.223	-28.120	-61.164	5.628	9.313	-1.811	-9.335	17.096	-26.432
unequal	is	10,000	0.117	12.754	-27.415	-53.028	6.614	9.102	0.360	-7.357	15.627	-22.984
unequal	is	15,000	0.228	12.743	-27.552	-51.677	6.690	9.035	0.665	-7.124	15.513	-22.637
equal	i	5,000	0.619	0.212	3.155	0.835	-1.585	-1.380	-0.381	0.211	1.167	-0.956
equal	i	10,000	1.489	0.114	-1.531	-0.364	0.033	-0.506	-0.130	-0.128	0.595	-0.723
equal	i	15,000	-0.334	-0.078	-3.951	-0.063	-1.526	-0.914	0.122	-0.964	0.998	-1.962
unequal	i	5,000	7.145	0.114	9.138	0.964	9.470	2.984	-0.241	4.225	4.294	-0.069
unequal	i	10,000	3.812	0.164	3.438	0.192	4.840	1.151	-0.222	1.911	1.974	-0.063
unequal	i	15,000	3.472	-0.035	-0.964	0.214	3.769	1.149	0.159	1.109	1.394	-0.286
equal	s	5,000	-0.040	-43.992	0.096	-80.516	-72.194	0.202	-132.074	-46.931	46.931	93.863
equal	s	10,000	0.077	-45.715	0.078	-79.923	-79.269	0.166	-131.588	-48.025	48.025	96.050
equal	s	15,000	0.017	-44.074	0.077	-79.861	-79.716	-0.010	-132.034	-47.943	47.943	95.886
unequal	s	5,000	0.054	-19.015	0.074	-81.074	-63.713	-0.245	-114.553	-39.782	39.782	79.563
unequal	s	10,000	0.042	-19.071	0.005	-78.567	-75.119	0.116	-114.635	-41.033	41.033	82.066
unequal	s	15,000	0.005	-19.195	-0.040	-77.018	-80.136	0.107	-115.412	-41.670	41.670	83.340

Table 12: Baseline Model – Proportional bias when sample is stratified with large differences in the probability of selection

Class proportion	Mean separation	Sample size	Proportional bias in estimated parameters							Ave. prop bias	|Ave prop bias|	Difference
			Intercept mean (class 1)	Slope mean (class 1)	Intercept variance	Slope variance	Mixing proportion	Intercept mean (class 2)	Slope mean (class 2)			
equal	is	5,000	-43.563	-35.077	-47.850	-61.085	-57.159	-35.685	-46.233	-46.664	46.664	-93.329
equal	is	10,000	-43.719	-35.075	-45.564	-57.949	-57.266	-35.775	-46.357	-45.958	45.958	-91.916
equal	is	15,000	-43.662	-34.987	-41.140	-52.909	-57.221	-35.598	-46.384	-44.557	44.557	-89.115
unequal	is	5,000	-22.435	-17.929	-45.944	-65.796	-23.792	-34.907	-44.669	-36.496	36.496	-72.992
unequal	is	10,000	-22.535	-17.923	-42.747	-61.917	-23.872	-35.015	-44.680	-35.527	35.527	-71.054
unequal	is	15,000	-22.510	-17.987	-42.354	-61.132	-23.855	-34.927	-44.699	-35.352	35.352	-70.704
equal	i	5,000	-39.641	0.179	-5.626	0.329	-50.858	-32.199	-0.571	-18.341	18.486	-36.827
equal	i	10,000	-41.340	-0.148	-8.747	-0.446	-53.892	-33.355	0.002	-19.704	19.704	-39.408
equal	i	15,000	-42.159	0.019	-9.687	0.249	-54.797	-33.670	-0.266	-20.044	20.121	-40.165
unequal	i	5,000	-17.077	0.026	-10.018	0.336	-16.907	-26.968	-0.130	-10.105	10.209	-20.314
unequal	i	10,000	-18.420	0.100	-13.726	0.120	-19.659	-28.866	-0.210	-11.523	11.586	-23.109
unequal	i	15,000	-19.325	-0.042	-13.315	-0.279	-20.896	-30.408	0.081	-12.026	12.050	-24.076
equal	s	5,000	0.093	-41.115	0.090	-59.174	-58.219	-0.018	-50.415	-29.822	29.875	-59.697
equal	s	10,000	-0.022	-46.479	0.124	-65.827	-65.420	-0.119	-55.113	-33.265	33.301	-66.566
equal	s	15,000	-0.068	-46.246	0.081	-65.308	-65.681	-0.122	-54.853	-33.171	33.194	-66.365
unequal	s	5,000	0.126	-19.286	0.049	-48.513	-67.987	-0.192	-44.961	-25.823	25.873	-51.697
unequal	s	10,000	-0.100	-21.665	0.170	-52.600	-74.131	-0.172	-48.977	-28.211	28.259	-56.470
unequal	s	15,000	-0.580	-25.700	0.220	-58.800	-79.924	12.634	-57.433	-29.941	33.613	-63.554

Based on the simulation results shown in Tables 10-12, the difference between the average bias and the absolute value of this value was much greater when the sample was disproportionately sampled, compared to a SRS sample. This indicates that there was much greater variability in parameter estimates. This variability, combined with the greater proportion of bias suggests that modeling adjustments may be necessary to accommodate data that are collected using complex sampling strategies. The results from three alternate modeling techniques (Experimental Models I, II, and III) are presented in Sections 4.2 - 4.4.

4.1.3 Standard Errors

Table 13 provides the standard errors for the Baseline Model, when simple random sampling was used to select cases. As expected, the standard errors get smaller as sample size increases. Furthermore, the standard error for the intercept variance parameter has the largest standard error, while the mean of the slope is the smallest standard error across all sample sizes.

Table 13: Baseline Model with SRS - standard errors of parameters, by sample size

Parameter	**n=5,000**	**n=10,000**	**n=15,000**
Mean Intercept	0.391	0.221	0.180
Mean Slope	0.194	0.153	0.133
Variance Intercept	0.766	0.509	0.381
Variance Slope	0.231	0.166	0.139

Next, Table 14 provides the standard errors for the Baseline Model when the sample selection process is more complex. With values similar to Table 13, the standard errors again decrease as the sample size increases. Here we can see that there is no clear difference between when the probabilities of selection are similar for each stratum (the 60-40 condition) and when the probabilities of selection vary greatly (the 80-20 condition).

Table 14: Baseline Model with complex sampling - standard errors of parameters, by sample size and probability of selection

Parameter	n=5,000		n=10,000		n=15,000	
	Strata 60-40	Strata 80-20	Strata 60-40	Strata 80-20	Strata 60-40	Strata 80-20
Mean Intercept	0.309	0.160	0.215	0.103	0.169	0.082
Mean Slope	0.070	0.087	0.081	0.054	0.028	0.043
Variance Intercept	1.192	0.961	0.735	0.805	0.499	0.743
Variance Slope	0.363	0.229	0.191	0.178	0.104	0.158

4.1.4 Entropy

Looking at the entropy results for the Baseline Model in Table 15, we see that individuals are categorized into latent classes with different levels of precision, depending on the method used to select the sample.

Table 15: Baseline Model – entropy by sample selection condition

Sample selection process	Entropy
SRS	0.485
Strata 60-40	0.619
Strata 80-20	0.701

Table 15 indicates that the SRS condition results in a very low proportion of individuals correctly categorized into the correct class. This entropy value is similar to that found in the literature; Lubke and Muthén, (2007) found entropy values ranging from .33 to .82 in their simulations of correctly specified mixture models.

Looking at entropy by the class proportion condition, the Baseline Model does only a slightly better job when the classes are of different sizes, as shown in Table 16.

Table 16: Baseline Model – entropy by class proportion condition

Class proportion	Entropy
Class 50-50	0.587
Class 70-30	0.617

Furthermore, when we look at entropy results by the class separation condition, we see that when both the intercept and slope means differ, the entropy values are higher than when only one factor mean differs. This is shown in Table 17.

Table 17: Baseline Model – entropy by mean separation condition

Mean separation	Entropy
Only I differs	0.482
Only S differs	0.647
Both I & S differ	0.676

Lastly, sample size did not seem to affect entropy, as shown in Table 18.

Table 18: Baseline Model – entropy by sample size condition

Sample size	Entropy
n=5,000	0.604
n=10,000	0.601
n=15,000	0.601

4.2 Experimental Model I

Experimental Model I represents the addition of weights to the Baseline Model. This section compares the results of this model to that of the Baseline Model to see if it performs better in terms of convergence, parameter bias, standard errors and entropy.

4.2.1 Model Convergence

While the Baseline Model had perfect convergence, one cell in Model I failed to reach the minimum number of replications needed to converge. This cell represented the smallest sample sizes tested within a single condition. This condition was where the probability of selection was 80-20, only the mean of the slope differed from class 1 to class 2, and the class proportions were 70-30. Additionally, two cells required additional replications, as shown in Table 19.

Table 19: The number of additional replications needed in Model I

Probability of selection	Mean separation	Class proportion	Sample size	Additional iterations needed
60-40	Slope only	70-30	5,000	8
60-40	Slope only	70-30	10,000	1

All converged solutions were admissible, with no negative estimates of variances.

4.2.2 Parameter Bias

The parameter bias estimates from Model I are presented in Tables 20 and 21. These tables are analogous to Tables 11 and 12 of the Baseline Model. Here, instead of an inappropriate application of the Baseline Model to data that have unequal probabilities of selection, a model that attempts to adjust for disproportionate sampling is used.

Table 20: Model I – Proportional bias when sample is stratified with small differences in the probability of selection

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-34.642	-45.009	-16.622	-28.931	44.610	-29.548	-19.347	-18.498	31.244	-49.743
equal	is	10,000	-34.702	-45.108	-16.355	-28.106	44.893	-29.732	-19.182	-18.327	31.154	-49.481
equal	is	15,000	-34.766	-45.068	-16.115	-28.366	44.930	-29.797	-19.199	-18.340	31.177	-49.517
unequal	is	5,000	-34.760	-44.862	-18.228	-31.471	2.813	-29.627	-19.448	-25.083	25.887	-50.970
unequal	is	10,000	-34.963	-45.310	-15.601	-27.894	8.178	-29.823	-19.730	-23.592	25.929	-49.521
unequal	is	15,000	-34.725	-45.090	-16.446	-27.704	8.173	-29.742	-19.256	-23.542	25.877	-49.418
equal	i	5,000	-25.382	-0.040	-9.883	0.043	13.925	-32.111	0.221	-7.604	11.658	-19.262
equal	i	10,000	-25.081	0.260	-16.414	0.362	7.734	-31.773	-0.381	-9.328	11.715	-21.043
equal	i	15,000	-25.974	-0.048	-16.750	0.350	9.797	-32.493	0.027	-9.299	12.206	-21.504
unequal	i	5,000	-28.292	0.117	-3.842	0.002	-2.149	-34.456	0.151	-9.781	9.858	-19.640
unequal	i	10,000	-26.007	0.089	-14.060	0.058	-0.791	-32.463	-0.231	-10.486	10.529	-21.015
unequal	i	15,000	-26.960	0.048	-14.978	0.166	-1.510	-33.031	-0.040	-10.901	10.962	-21.863
equal	s	5,000	-0.260	-25.417	0.062	-61.574	-5.900	0.532	-48.891	-20.207	20.377	-40.583
equal	s	10,000	0.057	-27.288	-0.179	-62.855	-7.295	-0.348	-49.435	-21.049	21.065	-42.114
equal	s	15,000	0.044	-27.775	-0.121	-63.145	-7.751	0.050	-49.782	-21.211	21.238	-42.450
unequal	s	5,000	-0.260	-25.417	0.062	-61.574	-5.900	0.532	-48.891	-20.207	20.377	-40.583
unequal	s	10,000	0.057	-27.288	-0.179	-62.855	-7.295	-0.348	-49.435	-21.049	21.065	-42.114
unequal	s	15,000	0.045	-27.765	-0.120	-63.137	-7.747	0.054	-49.782	-21.207	21.236	-42.443

Table 21: Model I – Proportional bias when sample is stratified with large differences in the probability of selection

Class proportion	Mean separation	Sample size	Proportional bias in estimated parameters							Ave. prop bias	|Ave prop bias|	Difference
			Intercept mean (class 1)	Slope mean (class 1)	Intercept variance	Slope variance	Mixing proportion	Intercept mean (class 2)	Slope mean (class 2)			
equal	is	5,000	6.595	22.654	-36.512	-50.393	-2.760	-34.687	-45.144	-20.035	28.392	-48.427
equal	is	10,000	6.631	22.724	-35.739	-50.514	-1.485	-34.685	-45.172	-19.748	28.136	-47.884
equal	is	15,000	6.782	22.501	-36.335	-51.151	-1.955	-34.621	-45.134	-19.988	28.354	-48.342
unequal	is	5,000	-35.272	-15.063	-75.706	-97.421	-16.018	-12.276	-16.901	-38.380	38.380	-76.759
unequal	is	10,000	-35.875	-15.379	-76.287	-98.556	-16.513	-11.738	-16.503	-38.693	38.693	-77.386
unequal	is	15,000	-36.278	-16.149	-76.560	-100.847	-17.049	-11.947	-16.922	-39.393	39.393	-78.787
equal	i	5,000	-1.722	0.189	-20.189	0.743	-13.383	-27.948	-0.149	-8.923	9.189	-18.112
equal	i	10,000	4.896	-0.297	-22.993	-0.300	-15.748	-30.924	-0.052	-9.345	10.744	-20.090
equal	i	15,000	9.387	0.042	-24.134	-0.392	-16.952	-32.828	0.255	-9.232	11.999	-21.230
unequal	i	5,000	-36.864	0.101	-58.287	0.323	-61.519	-7.246	-0.140	-23.376	23.497	-46.873
unequal	i	10,000	-37.867	0.060	-59.032	0.228	-62.743	-7.349	-0.236	-23.848	23.931	-47.779
unequal	i	15,000	-37.419	-0.008	-59.062	-0.217	-61.890	-6.669	0.080	-23.598	23.621	-47.219
equal	s	5,000	-0.314	23.152	0.284	-38.528	-20.254	-0.018	-45.792	-11.639	18.335	-29.973
equal	s	10,000	0.008	22.888	0.235	-38.349	-20.195	0.018	-45.641	-11.577	18.191	-29.767
equal	s	15,000	-0.062	23.742	0.002	-36.464	-19.716	-0.029	-45.179	-11.101	17.885	-28.986
unequal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	10,000	-0.069	-29.722	0.145	-68.804	-27.151	0.227	-29.407	-22.112	22.218	-44.330
unequal	s	15,000	0.041	-30.314	0.211	-69.737	-27.444	-0.092	-29.392	-22.390	22.461	-44.851

nc=non-converging cell

Tables 20 and 21 provide the proportional bias found in the parameters when the sample was selected with differing probabilities of selection. Table 20 shows the proportional bias when stratification resulted in a small difference in probability. Table 21 shows the same results for the large difference in probability condition. Unlike the Baseline model, the probability of selection did not seem to impact the number of cells with more than 5% bias for Model I. This is shown in Table 22, below.

Table 22: Model I compared to Baseline Model – Percentage of cells with more than 5% bias

Probability of selection	**Baseline Model**	**Model I**
Small difference (60-40)	50%	67%
Larger difference (80-20)	72%	68%

Furthermore, Table 23 displays the overall bias for Model I for each estimated parameter, compared to the Baseline Model. It provides a summary of the proportion of bias estimated for each of the parameters by type of sample selection process. Each cell in the table is an average across the sample size, factor means, and class proportion conditions. In contrast, Table 24 provides the same information by the mean separation condition (averaging across the sample size, sample selection process, and class proportion conditions).

Table 23: Model I compared to Baseline Model - Percent bias of model parameters by probability of selection

	Strata 60-40				Strata 80-20			
	Baseline*		Model I		Baseline*		Model I	
Parameter	Ave	|Ave.|	Ave	|Ave.|	Ave	|Ave.|	Ave	|Ave.|
Mean Intercept (class 1)	-12.630	14.527	-11.024	15.064	-20.941	20.965	-20.365	20.388
Mean Intercept (class 2)	3.489	3.829	-14.871	14.900	-21.409	22.813	-20.785	20.915
Mean Slope (class 1)	-11.844	16.103	1.831	14.411	-19.963	19.999	-23.943	24.000
Mean Slope (class 2)	-51.720	51.865	-22.437	22.476	-32.548	32.557	-22.924	22.968
Mixing Proportion	-32.493	36.609	-23.693	23.693	-48.419	48.419	7.706	12.855
Variance Intercept	-26.883	28.668	-34.115	34.218	-18.110	18.192	-9.765	9.779
Variance Slope	-52.053	52.298	-41.199	41.351	-39.483	39.598	-30.368	30.477

** SRS condition excluded*

Table 24: Model I compared to Baseline Model - Percent bias of model parameters by mean separation condition

	Means of I & S both differ				Only mean of I differs				Only mean of S differs			
	Baseline*		Model I		Baseline*		Model I		Baseline*		Model I	
Parameter	Ave	|Ave.|	Ave	|Ave.|	Ave	|Ave.|	Ave	|Ave.|	Ave	|Ave.|	Ave	|Ave.|
Mean Intercept (class 1)	-36.843	36.900	-24.665	27.999	-13.480	16.236	-21.441	23.821	-0.033	0.102	-0.065	0.111
Mean Intercept (class 2)	-12.660	22.658	-26.519	26.519	-15.249	16.129	-25.774	25.774	1.029	1.175	0.053	0.204
Mean Slope (class 1)	-15.134	21.420	-20.763	32.077	0.052	0.103	0.043	0.108	-32.629	32.629	-13.746	26.434
Mean Slope (class 2)	-38.583	38.754	-25.162	25.162	-0.149	0.210	-0.041	0.164	-87.671	87.671	-44.693	44.693
Mixing Proportion	-32.741	35.896	8.151	17.448	-16.834	19.853	-17.103	22.345	-71.792	71.792	-14.241	14.241
Variance Intercept	-63.256	63.256	-36.376	36.376	-4.319	6.941	-26.635	26.635	0.085	0.092	0.037	0.145
Variance Slope	-68.547	68.547	-51.779	51.779	0.174	0.366	0.114	0.265	-68.932	68.932	-57.002	57.002

** SRS condition excluded*

As can be seen from Table 23, Model I outperformed the Baseline Model on many of the estimated parameters, especially the mixing proportion parameter. However, there was no discernable pattern based on the probability of selection.

Looking at Table 24, Model I tends to exhibit less bias in the variances of the factors and the mixing parameter than the Baseline Model. In terms of the factor means, Model I offers some improvement over the Baseline Model when both the mean and the slope factor means differ, or when just the slope mean differs from class 1 to class 2.

4.2.3 Standard Errors

Table 25 shows that, for the variance parameters, there tends to be less variability in the estimated parameters when weights are added to the model, compared to the Baseline Model. However, for the intercept parameters, the estimates from the Baseline Model exhibit less variability. In all cases, the standard errors get smaller as sample size increases. Furthermore, the standard error for the intercept variance parameter is again the largest, while the variance of the slope is the smallest across all sample sizes.

Table 25: Model I compared to Baseline Model - Standard errors of parameters, by sample size condition

Parameter	**n=5,000**		**n=10,000**		**n=15,000**	
	Baseline*	**Model I**	**Baseline***	**Model I**	**Baseline***	**Model I**
Mean Intercept	0.234	0.301	0.159	0.203	0.126	0.137
Mean Slope	0.078	0.095	0.067	0.065	0.035	0.051
Variance Intercept	1.076	0.694	0.770	0.516	0.621	0.455
Variance Slope	0.296	0.194	0.185	0.141	0.049	0.118

** SRS condition excluded*

4.2.4 Entropy

Entropy values for experimental Model I are provided in Table 26. The entropy results for the weighted model indicate that this model does not classify individuals any more accurately than the Baseline Model. Moreover, the probabilities of selection do not seem to affect classification.

Table 26: Model I compared to Baseline Model – entropy by sample selection condition

Probability of selection	Baseline	Model I
Strata 60-40	0.619	0.638
Strata 80-20	0.701	0.630

When investigating entropy findings more closely, we also find that when the class proportions are manipulated to be equal, the weighted model classifies individuals more accurately than the Baseline Model. However, when class proportions are not equal, the opposite is true. This is shown in Table 27.

Table 27: Model I compared to Baseline Model – entropy by class proportion condition

Class proportion	Baseline*	Model I
Class 50-50	0.649	0.659
Class 70-30	0.672	0.608

** SRS condition excluded*

Similarly, when we look at entropy results by mean separation, we see that when both factor means are different, the entropy values are higher than when only one factor mean differs. This is shown in Table 28.

Table 28: Model I compared to Baseline Model – entropy by mean separation condition

Mean separation	Baseline*	Model I
Only I differs in means	0.513	0.546
Only S differs in means	0.777	0.569
Both I & S differ in means	0.691	0.751

** SRS condition excluded*

Again, sample size did not seem to affect entropy values, as shown in Table 29.

Table 29: Model I compared to Baseline Model – entropy by sample size

Sample size	Baseline*	Model I
n=5,000	0.657	0.635
n=10,000	0.660	0.634
n=15,000	0.664	0.634

** SRS condition excluded*

4.3 Experimental Model II

Experimental Model II represents the addition of a stratification covariate to the Baseline Model. This covariate models the sampling scheme. This section compares the performance of the model to that of the Baseline Model to see if the addition of a covariate adequately adjusts for a complex sample design.

4.3.1 Model Convergence

Although the Baseline Model did not have any problems, convergence was an issue for Model II. In fact, Model II had the worst convergence out of all of the models tested; 14 of the 36 cells did not attain 1,000 successful replications (a convergence rate of 61.1%). The non-converging cells are shown in Table 30.

Table 30: Model II: Non-converging cells

Probability of selection	Mean separation	Class proportion	Sample size
60-40	Slope only	70-30	5,000
60-40	Slope only	70-30	10,000
60-40	Slope only	70-30	15,000
60-40	Slope only	50-50	5,000
60-40	Slope only	50-50	10,000
60-40	Slope only	50-50	15,000
80-20	Intercept only	70-30	5,000
80-20	Intercept only	70-30	10,000
80-20	Intercept only	70-30	15,000
80-20	Intercept only	50-50	5,000
80-20	Intercept only	50-50	10,000
80-20	Intercept only	50-50	15,000
80-20	Slope only	70-30	5,000
80-20	Slope only	70-30	10,000

This convergence problem was also found by Lubke and Muthén (2007), in their study of mixture models. They found that mixture models with a dichotomous covariate (as in Model II) and with small mean separation, had lower convergence rates than the other models tested (72-84% converged).

However, for the remainder of the cells that did converge, no additional replications beyond the initial 1,000 were necessary. Furthermore, all converged solutions were admissible, with no negative estimates of variance.

4.3.2 Parameter Bias

The parameter bias estimates from Model II are presented in Tables 31 and 32. These tables are analogous to Tables 11 and 12 of the Baseline Model. Here, instead of an inappropriate application of the Baseline Model to data that have unequal probabilities of selection, Model II uses a stratification covariate to adjust for unequal sampling probabilities. However, it is important to note that the results are not representative of all manipulated conditions. This is because 38.9% of cells did not converge (as discussed in section 4.3.1), and as a result, parameter estimates were not obtained for these cells.

Table 31: Model II – Proportional bias when sample is stratified with small differences in the probability of selection

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-35.721	-9.970	-57.283	-69.884	-29.802	-6.371	-22.162	-33.028	33.028	-66.055
equal	is	10,000	-35.644	-10.260	-57.383	-69.838	-29.934	-6.321	-22.238	-33.088	33.088	-66.176
equal	is	15,000	-35.520	-10.170	-57.134	-70.456	-29.704	-6.359	-21.995	-33.048	33.048	-66.097
unequal	is	5,000	-18.792	-4.697	-52.972	-62.324	-13.313	-5.930	-24.369	-26.057	26.057	-52.113
unequal	is	10,000	-18.966	-4.595	-52.929	-61.702	-13.382	-6.009	-24.458	-26.006	26.006	-52.012
unequal	is	15,000	-18.862	-4.745	-52.916	-61.670	-13.440	-6.053	-24.599	-26.041	26.041	-52.081
equal	i	5,000	-0.731	0.064	-19.408	0.208	0.112	0.638	0.143	-2.711	3.043	-5.754
equal	i	10,000	-0.383	-0.048	-18.877	0.054	0.287	0.807	0.048	-2.588	2.929	-5.517
equal	i	15,000	-0.439	-0.143	-19.047	0.105	0.350	0.822	-0.016	-2.624	2.989	-5.613
unequal	i	5,000	-0.662	0.468	-19.539	0.672	-39.686	0.969	-1.308	-8.441	9.044	-17.484
unequal	i	10,000	-0.378	-0.049	-18.935	-0.003	-39.580	0.738	0.031	-8.311	8.530	-16.841
unequal	i	15,000	-0.258	0.058	-18.529	0.075	-39.450	0.825	-0.125	-8.201	8.474	-16.675
equal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	s	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	s	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc

nc= non-converging cells

Table 32: Model II – Proportional bias when sample is stratified with large differences in the probability of selection

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-75.107	-43.457	-71.496	-43.485	-61.799	-2.780	-25.436	-46.223	46.223	-92.446
equal	is	10,000	-74.982	-43.707	-71.402	-43.740	-62.065	-2.882	-25.221	-46.286	46.286	-92.571
equal	is	15,000	-75.433	-43.905	-71.635	-44.040	-62.305	-2.813	-25.377	-46.501	46.501	-93.002
unequal	is	5,000	-47.378	-34.835	-56.265	-45.445	-76.226	-4.579	-23.147	-41.125	41.125	-82.250
unequal	is	10,000	-47.672	-35.093	-57.097	-45.454	-76.630	-4.585	-23.724	-41.465	41.465	-82.930
unequal	is	15,000	-47.569	-35.091	-56.801	-45.268	-76.527	-4.499	-23.284	-41.291	41.291	-82.583
equal	i	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	i	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	i	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	i	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	i	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	i	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	s	5,000	-0.036	-60.239	0.101	-35.209	-79.801	0.807	-45.042	-31.345	31.605	-62.950
equal	s	10,000	0.021	-60.249	0.123	-35.148	-79.925	0.430	-44.367	-31.302	31.466	-62.768
equal	s	15,000	-0.056	-60.596	0.114	-35.475	-80.169	-0.417	-44.547	-31.592	31.625	-63.217
unequal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	15,000	-0.001	-46.091	0.087	-35.943	-34.708	0.141	-45.285	-23.114	23.179	-46.294

nc= non-converging cells

Tables 31 and 32 provide the proportional bias found in the parameters when the sample was stratified disproportionally. Table 31 shows the proportional bias when stratification resulted in a small difference in probability. Table 32 shows the same results for the large difference in probability condition. Compared to the Baseline Model, Model II has fewer cells with unacceptable levels of bias, when there are smaller differences in the probabilities of selection. However, when the differences in the probability of selection are small, the Baseline Model outperforms both experimental models. This is shown in Table 33, below.

Table 33: Model II compared to Baseline Model – Percentage of cells with more than 5% bias

Probability of selection	**Baseline Model**	**Model I**	**Model II**
Small difference (60-40)	50%	67%	57%
Larger difference (80-20)	72%	68%	74%

Furthermore, Table 34 displays the overall bias for Model II for each estimated parameter, compared to the Baseline Model discussed previously. It provides a summary of the proportion of bias estimated for each of the parameters by type of sample selection process. Each cell in the table is an average across the sample size, factor means, and class proportion conditions. In contrast, Table 35 provides the same information by the mean separation condition (averaging across the sample size, sample selection process, and class proportion conditions).

Table 34: Model II compared to Baseline Model - percent bias in model parameters by sample selection condition

	Strata 60-40				Strata 80-20			
	Baseline*		**Model II**		**Baseline***		**Model II**	
Parameter	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**
Mean Intercept (class 1)	-12.630	14.527	-36.821	36.826	-20.941	20.965	-13.863	13.863
Mean Intercept (class 2)	3.489	3.829	-2.118	2.393	-21.409	22.813	-2.687	3.487
Mean Slope (class 1)	-11.844	16.103	-46.326	46.326	-19.963	19.999	-3.674	3.772
Mean Slope (class 2)	-51.720	51.865	-32.543	32.543	-32.548	32.557	-11.754	11.791
Mixing Proportion	-32.493	36.609	-38.427	38.512	-48.419	48.419	-37.079	37.079
Variance Intercept	-26.883	28.668	-40.921	40.921	-18.110	18.192	-32.897	33.082
Variance Slope	-52.053	52.298	-69.015	69.015	-39.483	39.598	-20.629	20.753

** SRS condition excluded*

Table 35: Model II compared to Baseline Model - percent bias in model parameters by mean separation condition

	Means of I & S both differ				Only mean of I differs				Only mean of S differs			
	Baseline		**Model II**		**Baseline**		**Model II**		**Baseline**		**Model II**	
Parameter	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**
Mean Intercept (class 1)	-36.843	36.900	-44.304	44.304	-13.480	16.236	-0.475	0.475	-0.033	0.102	-0.018	0.029
Mean Intercept (class 2)	-12.660	22.658	-4.932	4.932	-15.249	16.129	0.800	0.800	1.029	1.175	0.240	0.449
Mean Slope (class 1)	-15.134	21.420	-23.377	23.377	0.052	0.103	0.058	0.138	-32.629	32.629	-56.794	56.794
Mean Slope (class 2)	-38.583	38.754	-23.834	23.834	-0.149	0.210	-0.205	0.278	-87.671	87.671	-44.810	44.810
Mixing Proportion	-32.741	35.896	-59.610	59.610	-16.834	19.853	-19.056	19.056	-71.792	71.792	0.106	0.106
Variance Intercept	-63.256	63.256	-55.275	55.275	-4.319	6.941	0.185	0.186	0.085	0.092	-35.444	35.444
Variance Slope	-68.547	68.547	-45.427	45.427	0.174	0.366	-19.661	19.911	-68.932	68.932	-68.651	68.651

** SRS condition excluded*

As can be seen from Table 34, Model II tends to outperform the Baseline Model when the there were small differences in the probabilities of selection. However, when the probability of selection varied greatly between individuals (the 80-20 condition), we see that the Baseline Model tends to provide more accurate parameter estimates than Model II.

In Table 35, we see that Model II sometimes outperforms the Baseline model when looking at the mean separation condition. However, it should be noted that for the 'slope only' condition, only 4 cells out of 12 converged. Similarly, for the 'intercept only' condition, only 6 cells converged.

4.3.3 Standard Errors

Table 36 shows that Model II generally has smaller standard errors than the Baseline Model.

Table 36: Model II compared to Baseline Model - Standard errors of parameters, by sample size and model

Parameter	**n=5,000**		**n=10,000**		**n=15,000**	
	Baseline*	**Model II**	**Baseline***	**Model II**	**Baseline***	**Model II**
Mean Intercept	0.234	0.114	0.159	0.080	0.126	0.060
Mean Slope	0.078	0.043	0.067	0.030	0.035	0.024
Variance Intercept	1.076	0.376	0.770	0.267	0.621	0.199
Variance Slope	0.296	0.086	0.185	0.060	0.049	0.050

** SRS condition excluded*

4.3.4 Entropy

Entropy values for the Baseline Model as compared to Model II are provided in Table 37. The covariate model does not assign individuals as accurately as the other models when there is a small difference in the probability of selection. However, with a larger difference in the probability of selection, Model II outperforms both.

Table 37: Model II compared to Baseline Model – entropy by probability of selection

Probability of selection	Baseline*	Model I	Model II
Strata 60-40	0.619	0.638	0.512
Strata 80-20	0.701	0.630	0.677

SRS condition excluded

Table 38 shows that when the class proportion condition is examined, Model II performs rather poorly compared to the previous two models.

Table 38: Model II compared to Baseline Model – entropy by class proportion condition

Class proportion	Baseline*	Model I	Model II
Class 50-50	0.649	0.659	0.560
Class 70-30	0.672	0.608	0.569

* *SRS condition excluded*

When we look at entropy results by mean separation, we see that the results for Model II are similar to the Baseline Model. This is illustrated in Table 39. We also see that classification is particularly poor for Model II when only the intercept of the mean differs from one class to the next. However, it should again be noted that 6 cells did not converge for this condition, and for the slope condition, only 8 of the 12 cells converged.

Table 39: Model II compared to Baseline Model – entropy by class separation condition

Mean separation	Baseline*	Model I	Model II
Only I differs	0.513	0.546	0.398
Only S differs	0.777	0.569	0.714
Both I & S differ	0.691	0.751	0.639

* *SRS condition excluded*

Again, sample size did not seem to greatly affect entropy values, as shown in Table 40.

Table 40: Model II compared to Baseline Model – entropy by sample size

Sample size	Baseline*	Model I	Model II
n=5,000	0.657	0.635	0.556
n=10,000	0.660	0.634	0.555
n=15,000	0.664	0.634	0.581

** SRS condition excluded*

4.4 Experimental Model III

Experimental Model III represents the addition of both weights and a covariate to the Baseline Model. This section compares the results of this model to the Baseline Model to see if it appropriately adjusts for a complex sample design.

4.4.1 Model Convergence

While the Baseline Model had 100% convergence, Model III experienced significant problems with convergence. Out of a total of 36 cells, 11 (31%) did not converge. The non-converging cells (cells that failed to attain 1,000 properly converged solutions) are shown in Table 41.

Table 41: Model III: Non-converging cells

Probability of selection	Mean separation	Class proportion	Sample size
60-40	Intercept only	70-30	5,000
60-40	Intercept only	70-30	10,000
60-40	Slope only	70-30	5,000
60-40	Slope only	70-30	10,000
60-40	Slope only	70-30	15,000
80-20	Slope only	50-50	5,000
80-20	Slope only	50-50	10,000
80-20	Slope only	50-50	15,000
80-20	Slope only	70-30	5,000
80-20	Slope only	70-30	10,000
80-20	Slope only	70-30	15,000

However, for the remainder of the cells that did converge, no additional replications beyond the initial 1,000 were necessary. All converged solutions were admissible, with no negative estimates of variances.

4.4.2 Parameter Bias

The parameter bias estimates from Model III are presented in Tables 42 and 43. These tables are analogous to Tables 11 and 12 of the Baseline Model. Here, instead of an inappropriate application of the Baseline Model to data that have unequal probabilities of selection, Model III tests the use of both weighting and a stratification covariate to adjust for unequal sampling probabilities. However, it is important to note that the results presented in Tables 42 and 43 are not representative of all manipulated conditions. This is because 31% of cells did not converge (as discussed in section 4.4.1), and therefore parameter estimates were not obtained for these cells.

Table 42: Model III– Proportional bias when sample is stratified with small differences in the probability of selection

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-29.893	-34.800	-19.486	-45.026	0.341	-43.390	-58.495	-32.964	33.062	-66.026
equal	is	10,000	-29.821	-34.753	-19.252	-44.972	-1.025	-43.155	-57.429	-32.915	32.915	-65.830
equal	is	15,000	-29.818	-34.650	-19.112	-44.957	-0.025	-42.961	-56.295	-32.546	32.546	-65.091
unequal	is	5,000	-10.147	-33.923	-2.879	-43.536	-18.481	-40.200	-59.152	-29.760	29.760	-59.519
unequal	is	10,000	-10.104	-33.921	-2.934	-43.625	-18.501	-38.177	-58.285	-29.364	29.364	-58.727
unequal	is	15,000	-10.064	-33.788	-2.838	-43.579	-18.459	-38.000	-57.012	-29.106	29.106	-58.212
equal	i	5,000	-25.185	-32.147	0.233	-0.070	-7.079	-23.380	1.165	-12.352	12.751	-25.103
equal	i	10,000	-25.861	-32.427	-0.061	-0.259	-10.370	-27.397	-0.262	-13.805	13.805	-27.610
equal	i	15,000	-25.829	-32.371	-0.038	-0.179	-10.031	-26.787	0.494	-13.535	13.676	-27.210
unequal	i	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	i	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	i	15,000	-6.488	-30.205	0.115	-0.441	-15.788	-17.129	0.795	-9.877	10.137	-20.015
equal	s	5,000	0.132	0.199	-25.068	-48.144	-47.702	-0.143	-60.540	-25.895	25.990	-51.885
equal	s	10,000	-0.009	0.090	-26.465	-48.856	-49.608	0.172	-61.875	-26.650	26.725	-53.375
equal	s	15,000	-0.073	0.142	-26.381	-48.745	-49.685	-0.024	-61.404	-26.596	26.636	-53.232
unequal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc

nc=non-converging cells

Table 43: Model III – Proportional bias when sample is stratified with large differences in the probability of selection

Class proportion	**Mean separation**	**Sample size**	**Proportional bias in estimated parameters**							**Ave. prop bias**	**|Ave prop bias|**	**Difference**
			Intercept mean (class 1)	**Slope mean (class 1)**	**Intercept variance**	**Slope variance**	**Mixing proportion**	**Intercept mean (class 2)**	**Slope mean (class 2)**			
equal	is	5,000	-58.399	-37.617	-51.679	-48.961	-40.075	-51.804	-56.938	-49.353	49.353	-98.707
equal	is	10,000	-58.367	-37.469	-51.691	-48.719	-42.206	-51.578	-57.986	-49.716	49.716	-99.433
equal	is	15,000	-58.320	-37.448	-51.671	-48.848	-36.755	-49.576	-55.581	-48.314	48.314	-96.628
unequal	is	5,000	-36.670	-37.088	-35.405	-47.028	-30.605	-46.103	-59.842	-41.820	41.820	-83.640
unequal	is	10,000	-36.800	-36.924	-35.477	-46.900	-30.590	-48.178	-62.160	-42.433	42.433	-84.865
unequal	is	15,000	-36.652	-36.927	-35.457	-46.723	-30.584	-46.836	-60.417	-41.942	41.942	-83.884
equal	i	5,000	-49.555	-37.740	0.108	-0.075	-1.001	-21.622	-0.452	-15.762	15.793	-31.555
equal	i	10,000	-49.457	-37.777	0.145	0.172	-2.113	-22.268	-0.408	-15.958	16.049	-32.007
equal	i	15,000	-50.348	-37.883	0.102	-0.066	-2.984	-24.060	0.123	-16.445	16.509	-32.955
unequal	i	5,000	-28.837	-36.668	-0.175	0.419	-1.412	-12.617	-0.235	-11.361	11.481	-22.841
unequal	i	10,000	-29.987	-37.415	0.038	-0.145	-2.315	-14.227	0.282	-11.967	12.059	-24.026
unequal	i	15,000	-30.131	-37.237	0.144	-0.197	-2.252	-15.016	-0.170	-12.123	12.164	-24.287
equal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	s	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
equal	s	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	5,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	10,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc
unequal	s	15,000	nc	nc	nc	nc	nc	nc	nc	nc	nc	nc

nc=non-converging cells

Tables 42 and 43 provide the proportional bias found in the parameters when the sample was selected with different probabilities of selection. Table 42 shows the proportional bias when stratification resulted in a small difference in probability. Table 43 shows the same results for the large difference in probability condition. Compared to the Baseline model, Model III does not provide a substantial improvement in the number of cells with more than 5% bias. This is shown in Table 44, below.

Table 44: Model II compared to Baseline Model – Percentage of cells with more than 5% bias

Probability of selection	Baseline Model	Model I	Model II	Model III
Small difference (60-40)	50%	67%	57%	70%
Larger difference (80-20)	72%	68%	74%	71%

Furthermore, Table 45 displays the overall bias for Model III for each estimated parameter, compared to the Baseline Model. It provides a summary of the proportion of bias estimated for each of the parameters by type of sample selection process. Each cell in the table is an average across the sample size, factor means, and class proportion conditions. In contrast, Table 46 provides the same information by the mean separation condition (averaging across the sample size, sample selection process, and class proportion conditions).

Table 45: Model III compared to Baseline Model - percent bias in model parameters by probability of selection

	Strata 60-40				Strata 80-20			
	Baseline*		Model III		Baseline*		Model III	
Parameter	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**
Mean Intercept (class 1)	-12.630	14.527	-43.627	43.627	-20.941	20.965	-15.628	15.648
Mean Intercept (class 2)	3.489	3.829	-37.349	37.349	-21.409	22.813	-25.581	25.647
Mean Slope (class 1)	-11.844	16.103	-21.752	21.841	-19.963	19.999	-11.090	11.143
Mean Slope (class 2)	-51.720	51.865	-23.922	24.021	-32.548	32.557	-31.722	31.722
Mixing Proportion	-32.493	36.609	-33.657	33.657	-48.419	48.419	-26.198	26.224
Variance Intercept	-26.883	28.668	-29.482	29.550	-18.110	18.192	-40.638	41.016
Variance Slope	-52.053	52.298	-18.574	18.574	-39.483	39.598	-18.955	19.007

** SRS condition excluded*

Table 46: Model III compared to Baseline Model - percent bias in model parameters by mean separation condition

	Means of I & S both differ				Only mean of I differs				Only mean of S differs			
	Baseline*		Model III		Baseline*		Model III		Baseline*		Model III	
Parameter	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**	**Ave**	**|Ave.|**
Mean Intercept (class 1)	-36.843	36.900	-33.755	33.755	-13.480	16.236	-32.168	32.168	-0.033	0.102	0.017	0.071
Mean Intercept (class 2)	-12.660	22.658	-35.776	35.776	-15.249	16.129	-35.187	35.187	1.029	1.175	0.144	0.144
Mean Slope (class 1)	-15.134	21.420	-27.323	27.323	0.052	0.103	0.061	0.116	-32.629	32.629	-25.971	25.971
Mean Slope (class 2)	-38.583	38.754	-46.073	46.073	-0.149	0.210	-0.084	0.202	-87.671	87.671	-48.582	48.582
Mixing Proportion	-32.741	35.896	-44.996	44.996	-16.834	19.853	-20.450	20.450	-71.792	71.792	0.002	0.113
Variance Intercept	-63.256	63.256	-58.299	58.299	-4.319	6.941	0.133	0.439	0.085	0.092	-61.273	61.273
Variance Slope	-68.547	68.547	-22.247	22.304	0.174	0.366	-5.535	5.535	-68.932	68.932	-48.999	48.999

** SRS condition excluded*

Like Model II, Model III tends to outperform the Baseline Model when the there were small differences in the probabilities of selection. However, when the probability of selection varied greatly between strata (the 80-20 condition), we see the Baseline Model tends to provide more accurate parameter estimates than Model III.

Table 46 shows that when only the slope mean differs between classes, Model III tends to outperform the Baseline Model. However, when only the mean of the intercept factor differs or when both the intercept and slope means differ, the results are less predictable.

4.4.3 Standard Errors

Table 47 shows that the model using both covariates and weighting generally has less variability in estimating parameters than the Baseline Model.

Table 47: Model III compared to Baseline Model - standard errors of parameters, by sample size and model

Parameter	**n=5,000**		**n=10,000**		**n=15,000**	
	Baseline*	**Model III**	**Baseline***	**Model III**	**Baseline***	**Model III**
Mean Intercept	0.234	0.160	0.159	0.117	0.126	0.092
Mean Slope	0.078	0.054	0.067	0.038	0.035	0.029
Variance Intercept	1.076	1.009	0.770	0.902	0.621	0.860
Variance Slope	0.296	0.198	0.185	0.158	0.049	0.142

** SRS condition excluded*

4.4.4 Entropy

Entropy values for the Model III as compared to the other experimental models and the Baseline Model are provided in Table 48. While entropy values are low across the board, we see that overall, Model III is the most accurate in classifying individuals into their latent classes. It is the most accurate when the probabilities of selection vary the most (the 80-20 condition).

Table 48: Model III compared to Baseline Model – entropy by probability of selection

Sample selection	Baseline*	Model I	Model II	Model III
Strata 60-40	0.619	0.638	0.512	0.688
Strata 80-20	0.701	0.630	0.677	0.706

** SRS condition excluded*

Entropy values for Model III are higher when the class proportions are manipulated to be unequal, as shown in Table 49.

Table 49: Model III compared to Baseline Model – entropy by class proportion

Class proportion	Baseline*	Model I	Model II	Model III
Class 50-50	0.649	0.659	0.560	0.648
Class 70-30	0.672	0.608	0.569	0.769

** SRS condition excluded*

When we look at entropy results by mean separation, we see that when both the intercept mean and slope mean are different in each class, the entropy values tend to be higher than when only one factor mean differs, as shown in Table 50. This is similar to Model I and the Baseline Model.

Table 50: Model III compared to Baseline Model – entropy by mean separation condition

Mean separation	Baseline*	Model I	Model II	Model III
I only	0.513	0.546	0.398	0.560
S only	0.691	0.569	0.714	0.560
I and S	0.777	0.751	0.639	0.844

** SRS condition excluded*

Again, sample size did not seem to affect entropy values, as shown in Table 51.

Table 51: Model III compared to Baseline Model – entropy by sample size condition

Sample size	Baseline*	Model I	Model II	Model III
n=5,000	0.657	0.635	0.556	0.699
n=10,000	0.660	0.634	0.555	0.699
n=15,000	0.664	0.634	0.581	0.693

** SRS condition excluded*

4.5 Decomposition of Results

Additional explorations of the data were conducted to decompose the across-cell variability in the performance criteria. This was conducted in SPSS using the GLM univariate analysis procedure. Since the data contain empty cells, the test hypothesis for Type III sums of squares was not suitable (SPSS, 1999). Instead, Type IV sums of squares was used.

The effect of the experimental conditions on each of the performance criteria (model convergence, parameter bias, standard errors, and entropy) was tested. That is, whether the means of the groups formed by crossing each experimental condition have performance criteria that are different enough to have formed by chance. As all of the conditions were included in the analysis as separate factors, significance tests could not be calculated due to the lack of within group variation. Instead, the proportion of variation accounted for was reviewed, and any factor that accounted for 10% of more of the variance is reported.

4.5.1 Iterations

In decomposing model convergence, the number of additional replications over the initial 1,000 necessary for the model to converge was explored. While no main effect accounted for more than 10% of variance, one interaction effect reached this threshold. This was a three-way interaction of sample size, sample selection, and class separation, and it accounted for 12.5% of the variance. However, a three-way interaction can be difficult to interpret. It basically deals with the question: Are two-way interaction differences between the population means of 2 explanatory variables different for the various levels of the 3rd explanatory variable? In this instance, the effects of sample size and mean separation on the number of replications differed across levels of probability of selection. This 3-way interaction appears to be driven by the cells where the mean separation of the slope is combined with small differences in the probability of selection, which is then modified by the 3 levels of sample size.

4.5.2 Parameter bias

Table 52 shows that when estimating model parameters, the sample size, the sample selection process, and the latent class proportion do not substantially explain the variability in the parameter estimates (that is, they did not account for 10% or more of the variance). Instead, mean separation is an important predictor of bias for most of the parameters. The model used in the analysis (i.e. Baseline Model and Experimental Models I – III) also plays an important role in accounting for the variance in three of the parameters.

Table 52: Main effects –Proportion of variance explained in the bias of estimated parameters

	Manipulated Conditions				
Parameter	**Sample size**	**Model**	**Sample selection**	**Mean separation**	**Class proportion**
Mean Intercept (C1)				33.7%	
Mean Slope (C1)		14.6%			
Mean Intercept (C2)		28.1%		14.1%	
Mean Slope (C2)				54.9%	
Variance Intercept				42.3%	
Variance Slope				77.1%	
Mixing Proportion		24.0%			

In terms of the interactions among the tested conditions, those that account for 10% or more of the variance in proportional bias are shown in Table 53.

Table 53: Interaction effects – Proportion of variance explained in the bias of estimated parameters

Parameter	Manipulated Conditions		
	Model x Sample selection	Model*Mean separation	Model x Class proportion
Mean Intercept (C1)	12.4%		11.2%
Mean Slope (C1)	20.1%	13.3%	
Mean Intercept (C2)		33.9%	
Mean Slope (C2)			
Variance Intercept		10.1%	
Variance Slope			
Mixing Proportion	12.9%	15.7%	

Of the 26 possible interactions, only three accounted for more than 10% of the variance in the parameter bias, as shown in Table 53. All of these interactions involve the type of model. The main effect of model is modified by how the sample is selected, by mean separation, and by class proportion.

4.5.3 Standard errors

Neither sample size, the sample selection procedure, nor the class proportion condition has a substantial effect on the standard errors associated with the estimated parameters. However, mean separation explained 10% or more of the variance in all but one case. Furthermore, the model used to estimate parameters explained the standard errors of the variance parameters and the mixing parameter, as shown in Table 54.

Table 54: Main effects – Standard errors of estimated parameters

Parameter	Manipulated Conditions				
	Sample size	Model	Sample selection	Mean separation	Class proportion
Mean Intercept (C1)				35.1%	
Mean Slope (C1)				35.7%	
Mean Intercept (C2)				49.5%	
Mean Slope (C2)				40.6%	
Variance Intercept		28.3%		37.6%	
Variance Slope		18.9%		21.7%	
Mixing Proportion		16.3%			

As might be expected after looking at the main effects, the interaction among the model and separation conditions explained more than 10% of the variance for three of the parameters.

Table 55: Interaction effects – Standard errors of estimated parameters

Parameter	Manipulated Conditions		
	Sample selection*Mean separation	Model*Mean separation	Model*Class proportion
Mean Intercept (C1)			
Mean Slope (C1)			
Mean Intercept (C2)		11.1%	
Mean Slope (C2)	15.3%		
Variance Intercept			
Variance Slope		10.2%	
Mixing Proportion		15.7%	22.8%

4.5.4 Entropy

Entropy is a summary measure that gauges to accuracy of classification of individuals to their correct latent class, as discussed in Section 2.2.3.1. The variability of this measure can largely be explained by the model used and mean separation, as shown in Table 56. None of the other manipulated conditions accounted for more than 10% of the variability of the entropy parameter.

Table 56: Main effects – entropy

	Manipulated Conditions				
Parameter	**Sample size**	**Model**	**Sample selection**	**Mean separation**	**Class proportion**
Entropy		13.9%		39.1%	

Furthermore, the interaction effect between mean separation and sample selection criteria accounted for 11.3% of variance. That is, the effect of the degree of mean separation is modified by the procedure used to select the sample. No other interaction effect reached the 10% threshold.

4.6 Summary of Results

At this point it may be necessary to summarize the main effects presented in the previous sections. Table 57 provides this summary in table form.

Table 57: Performance criteria – Summary of main effects

	Manipulated Conditions				
	Sample size	**Model**	**Sample selection**	**Mean separation**	**Class proportion**
Iterations	Not affected	Not affected	Not affected	Not affected	Not affected
Parameter bias	Not affected	Some parameters affected (3 out of 7)	Not affected	Most parameters affected (5 out of 7)	Not affected
Standard errors	Not affected	Some parameters affected (3 out of 7)	Not affected	Most parameters affected (6 out of 7)	Not affected
Class assignment	Not affected	Affected	Not affected	Affected	Not affected

As part of the decompositions analysis, the number of additional iterations needed to obtain 1,000 properly converged solutions were analyzed (non-converging cells were excluded from this analysis.) None of the manipulated conditions played an important role in the number of additional iterations. However, the interaction of sample size, mean separation, and the probability of selection was important. Thus, if convergence is a concern, a combination of larger sample size, mean separation of both factors, and a smaller difference in the probability of selection would all help.

In estimating model parameters, mean separation and model play important roles. The same is true for the estimation of standard errors and the assignment of individuals into their correct latent class. The bottom line is that when looking at these four performance criteria (iterations, parameter bias, standard errors, and class assignment) mean separation is the most important aspect in GMM, followed by the type of model modification used (Baseline, and Experimental Models I-III).

4.7 Limitations

As with all simulation studies, a number of limitations apply to this research. Relatively few comprehensive GMM simulations have been conducted to date. Therefore, every effort was made to formulate a population model that was reasonably representative of first-stage GMM applications that appear in the literature. Nonetheless, as with any study with simulated data, there are almost an infinite number of conditions that could have been tested. For example, not all methods of adjusting for unequal probabilities of selection were tested in this research study. It is hoped that future research in this area will explore the use of a weighted covariance matrix (using other software than Mplus, as Mplus does not allow the use of correlation matrices when the MIXED command is employed).

Furthermore, the current research was limited to one statistic to gauge the accuracy of classifications to latent classes. While the entropy measure is provided as standard output in the Mplus software, there are significant concerns about its accuracy (discussed further in Section 5.1). Also, this research did not utilize observed variables with non-normal distributions. Such variables are common in survey data as survey research often measures behaviors in which only a minority of the population is engaged. Therefore, it is hoped that future research will include non-normally distributed variables. In addition, only unequal probabilities resulting from stratification in the sample design were explored in this research. In most nationally representative data sets both stratification and clustering are employed in the sample design. Therefore, it will likely be necessary not only to account for unequal probabilities of selection, but also for clustering via multilevel modeling or adjustments to the standard errors using Jacknife or Bootstrapping methodology.

This research also assumed that the true number of latent classes was known. While this is a common assumption in simulation studies (e.g. Gagné, 2004; Mann, 2009; Lubke & Muthén, 2007) it ignores model selection decisions that have to be made by applied researchers. Therefore, the current research ignores the possibility of extracting spurious latent classes.

Lastly, high rates of non-convergence for certain models may have impacted the generalizability of these findings. Using only the replications that resulted in proper solutions means that the results should be considered the upper bounds. In applied settings, steps may be taken after a model fails to converge, and these steps were not taken in this current research.

CHAPTER 5: DISCUSSION

The purpose of this research is to determine if complex sampling procedures affect the results of standard growth mixture models, and if they do, how these mixture models could be improved to better accommodate such sampling. Specifically, there are three main research goals of this research--they are to:

1. Determine if, and under what circumstances, ignoring unequal sampling probabilities would result in biased parameter estimates and poor class assignments;
2. Develop and test methods for incorporating unequal sampling probabilities into GMM; and
3. Recommend the most practical and efficient procedure given the constraints of the field and available software.

This discussion chapter will attempt to address each of these research goals in turn, using the results of the Monte Carlo simulation presented in the previous section.

5.1 Consequences of Ignoring Unequal Sampling Probabilities

To determine the consequences of ignoring unequal sampling probabilities, it is first necessary to examine the performance of the Baseline Model when the sample is collected via SRS. This represents an appropriate application of standard GMM. The next step is to compare these data to when the data are collected with varying probabilities of selection. This comparison should provide the consequences (if any) of ignoring sample design when using a standard GMM.

In terms of model convergence, there was not a substantial difference in the Baseline Model when data were from a SRS or sampled with different probabilities; both had convergence rates of 100%. Additionally, the SRS run resulted in 8 cells that required additional replications (over the initial 1,000), whereas the runs using differing probabilities of selection resulted in 6 cells requiring additional replications. These convergence results are not too surprising. Although mixture models in general are prone to local maxima of the likelihood (Bauer & Curran, 2003; Lubke et al. 2002; Muthén, 2001a), the degree to which convergence is likely to be a problem is related to the complexity of the model. As this research focused on very simple models, high convergence rates were expected.

In terms of biased parameters, the Baseline Model provides parameters estimates with only 1.45% bias when the sample is selected via SRS. When a more complex sampling scheme is introduced (either when the sample is stratified with small or large differences in the probability of selection), the overall proportional bias increases to nearly 30%, as shown in Table 58.

Table 58: Baseline Model - Proportional bias* across 3 sampling conditions.

Parameters	SRS	60-40	80-20
Intercept mean (class 1)	0.70	14.53	20.97
Slope mean (class 1)	1.98	16.10	20.00
Intercept variance	2.24	28.67	18.19
Slope variance	2.25	52.30	39.60
Mixing proportion	2.14	36.61	48.42
Intercept mean (class 2)	0.28	3.83	22.81
Slope mean (class 2)	0.60	51.87	32.56
Average over all parameters	*1.45*	*29.13*	*28.93*

** Absolute values*

Here, we can see that when GMM is applied to SRS data, the resulting parameter estimates are reasonably accurate. However, once the sampling strategy becomes even slightly more complex (when the probabilities of selection are similar across strata) parameter estimates are no longer reliable. These results are similar to those found by DuMouchel and Duncan (1983) and Nathan and Holt (1980) in their investigation of regression models. They found that, without any adjustment for unequal probabilities of selection, parameter estimates were biased. These findings are also similar to others examining statistical techniques such as ANOVA (e.g., Potthoff et al., 1992; Wedel et al., 1977), single-level (e.g., Hahs, 2003; Kaplan & Ferguson, 1999; Stapleton, 2006a; Stapleton, 2006b) and multilevel SEM (e.g., Lohr, 1999; Stapleton, 2002), growth models (Asparouhov, 2005), and latent class analysis (e.g., Patterson et al., 2002). This body of research has determined that if unequal selection probabilities are not accounted for, the resulting population estimates may exhibit substantial bias.

In terms of standard errors, there was not much change moving from SRS to a sampling design with unequal probabilities of selection, as shown in Table 59. This finding is also consistent with much of the previous research in this area.

While clustering of data tend to result in accurate parameter estimates, but inflated standard errors, having unequal probabilities of selection tends to have the opposite effect—biased parameter estimates but accurate standard errors (Skinner et al., 1989).

Table 59: Baseline Model – Standard errors across 3 different sampling conditions.

Parameters	**SRS**	**60-40**	**80-20**
Intercept mean (class 1)	0.18	0.23	0.11
Slope mean (class 1)	0.13	0.06	0.06
Intercept variance	0.38	0.81	0.84
Slope variance	0.14	0.22	0.19
Mixing proportion	0.08	0.10	0.06
Intercept mean (class 2)	0.26	0.34	0.38
Slope mean (class 2)	0.20	0.20	0.24
Average over all parameters	0.20	0.28	0.27

Lastly, the classification of individuals to their correct latent class was examined. Lubke and Muthén (2007) in their research on mixture models determined that entropy values below .60 are generally related to misclassifying approximately 20% or more of subjects. Entropy values around .80 or above are related to 90% correct class assignment. However, Mann (2009), in her simulation study of one and two factor mixture models, used the entropy measure to predict correct class assignment in a simple linear regression. She found that for over 80% of her cells, the regression coefficients were negative, indicating an inverse relationship between correct class assignment and entropy.

Table 60 shows that individuals were classified more accurately when the sampling design was ignored. This finding runs counter to expectations, although Lubke and Muthén (2007) also found that classification error rates in assigning subjects to their correct class were generally quite high and variable.

Table 60: Baseline Model – entropy values across 3 different sampling conditions.

	SRS	60-40	80-20
Entropy	.48	.62	.70

The estimation of GMM under the SRS condition, compared to the differential probability condition, yielded some expected results. Namely, convergence is high in both cases and standard errors were also consistent across both cases. Furthermore, parameters in the SRS condition were estimated with minimal bias, whereas when subjects have different probabilities of selection, bias in the parameter estimates rose to unacceptable levels; close to 30%. Some of the results were more unexpected. The Baseline Model simulations indicate that GMM do not classify individuals accurately, even when data are collected with equal probabilities of selection (SRS). Entropy does not meet the .80 threshold discussed previously, and improves as the probabilities of selection are more unequal.

5.2 Incorporating Unequal Sampling Probabilities into GMM

When the Baseline Model with data collected via SRS was compared to the Baseline Model where the data are collected with unequal probabilities of selection, it was demonstrated that the addition of a complex sampling structure did not substantially affect convergence rates, or standard errors. However, the application of an unadjusted GMM to data with unequal probabilities of selection resulted in substantial bias in parameter estimates. Therefore, it is important to identify methods that would allow GMM to estimate parameters more accurately under these conditions.

To identify these methods, it is useful to return to the results of the decomposition analysis presented in Section 4.5. Here, we found that two of the manipulated conditions accounted for substantial amounts of variability in parameter estimates: mean separation and experimental model used. When the mean of one factor differed between classes, the parameter estimates associated with the opposite mean were estimated more accurately. For example, when the mean of the intercept factor differs from class 1 to class 2, the mean slope and slope variance parameters tend to be estimated with more accuracy. This finding is generally consistent with previous research which has found that class separation is a major factor influencing the performance of mixture models. Lubke and Muthén (2007), found that in general, classes have to be separated at least by a multivariate Mahalanobis distance of 1.5 or greater to successfully estimate parameters (proportions of recovered parameters above .90). However, in an empirical study, mean separation is not usually under the control of the researcher.

More importantly, from a practical perspective, the decomposition analysis also found that the type of model used explained substantial amounts of variance in the parameter estimates. Of the experimental models, Model I produces the best results, although the percentage of parameter bias is still substantially more than when the Baseline Model is correctly applied to data sampled via SRS, as shown in Table 61. This implies that, while adding weights to the Baseline Model may decrease the amount of bias slightly, the parameter estimates still cannot be trusted.

Table 61: Average proportional parameter bias* across all tested models.

Parameters	**Baseline Model**		**Model I**	**Model II**	**Model III**
	SRS	**Sample**			
Intercept mean (class 1)	0.70	17.75	17.80	24.30	29.08
Slope mean (class 1)	1.98	18.05	19.34	23.12	16.28
Intercept variance	2.24	23.43	21.65	37.73	29.79
Slope variance	2.25	45.95	35.76	36.65	35.51
Mixing proportion	2.14	42.51	18.12	42.69	18.80
Intercept mean (class 2)	0.28	13.28	17.99	2.99	31.26
Slope mean (class 2)	0.60	42.21	22.73	21.24	28.03
Average over all parameters	*1.45*	*29.03*	*21.91*	*26.96*	*26.96*

**Absolute values*

It is unclear whether these finding are specific to the population model tested. It seems reasonable to expect that the pattern of bias observed in any given situation might be somewhat idiosyncratic and related to the exact characteristics of the population model. Regardless, these results clearly demonstrate that all GMM adjustments tested (Experimental Models I – III) suffer from a substantial amount of parameter bias under the conditions set forth in this study. Of the Experimental Models, Models II and III performed the worst. However, even for Model I, the percentage of bias in parameter estimates is still unacceptably high.

It was hypothesized that, even though the sample sizes tested were considerably larger than presently found in the applied literature, they may still be too small to accurately estimate parameters. Five additional runs were conducted at a sample size of 500,000 to test this hypothesis. These additional runs resulted in parameter estimates that were similar to the original estimates presented earlier. Therefore, it is unlikely under these conditions, that an increase in sample size above those already tested will improve GMM parameter estimates.

These results are disappointing. It was hoped that a simple strategy to adjust GMM for disproportionate sampling could be developed to reduce bias in parameter estimates to acceptable levels. This was clearly not the case.

In terms of convergence, the Baseline Model and Model I both had excellent rates of convergence (100% and 91.7% respectively). However, Models II and III had very poor rates of convergence (61.1% and 69.4% respectively). In mixture models, convergence is determined not only by the derivatives of the log-likelihood but also by the absolute and relative changes in the log-likelihood and the changes in class counts. Thus, the computational load of GMM estimation is extremely heavy. As models become more complex, they become more difficult to estimate as evidenced by failed M step iterations, a nonpositive Fisher information matrix, or other computational problems (Wang & Bodner, 2007).

Although the decomposition analysis did not indicate that sample size accounted for a substantial amount of variability in the number of iterations needed to attain 1,000 properly converged replications, it was hypothesized that a dramatic five-fold increase in sample size might improve the convergence rates of Models II and III. Therefore, five additional runs of these two experimental models were conducted at a sample size of 500,000. All of these previously non-converging cells converged at this sample size. Despite the fact that these additional runs indicate that a much larger sample size will improve convergence, it is important to note that the resulting parameter estimates still evidenced a substantial degree of bias.

In an applied setting, a useful way to avoid convergence problems due to poor starting values is to build up a model by estimating the model parts separately to obtain appropriate starting values for the full model. Furthermore, some researchers (e.g., Greenbaum et al., 2005; Muthén, 2004) have recommended specifying the growth shape in each latent class according to a priori theoretical hypotheses to reduce the computational load.

Entropy values from this study were generally poor, with an average of .65 (ranging from .38 - .90). This range is similar to that of Mann (2009), whose entropy values ranged from .33 to .70. Model II performed very poorly in terms of entropy, with no cell reaching the .80 threshold. However, Models I and III performed better, with 8 and 10 cells (respectively) making the cutoff. In general, Models I and III obtained satisfactory results if both the intercept and slope means differed from class 1 to class 2. The Baseline Model also resulted in 12 cells with entropy values of .80 or more. In this case, the highest values were found when the stratification resulted in a large difference in probability and only the slope mean differed between classes.

It is possible that Entropy may not be an adequate measure of correct class assignment. Mann (2009), in her simulation study of one and two factor mixture models, used the entropy measure to predict correct class assignment in a simple linear regression. She found that for over 80% of her cells, the regression coefficients were negative, indicating an inverse relationship between correct class assignment and entropy. For the remaining cells with positive regression coefficients “little of the variation in correct class assignment could be explained by entropy” (p. 99).

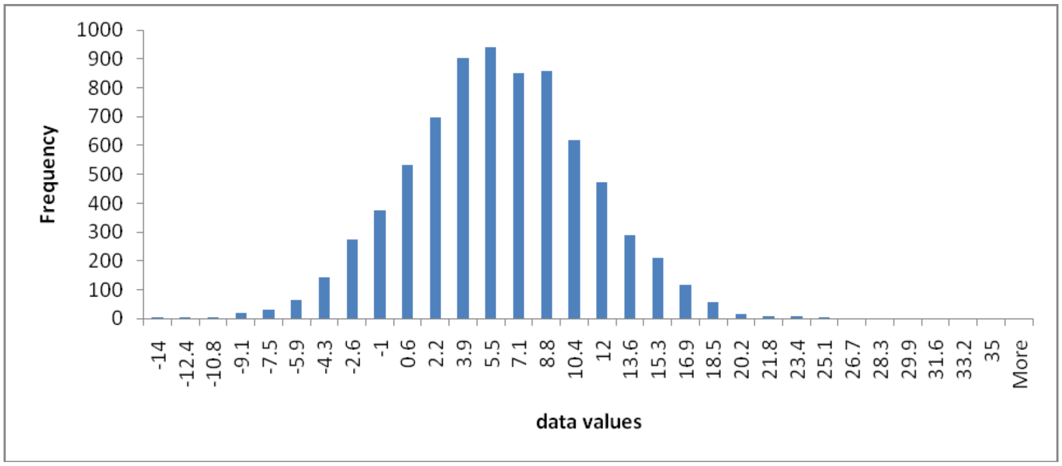
5.3 Recommendations

This research has found that a model-based approach (Experimental Models II and III) does not improve the accuracy of parameter estimates when individuals are sampled with disproportionate sampling probabilities. Therefore, the use of the stratification option in Mplus to adjust for a stratified sample is not recommended. Not only does this method often fail to converge, when it did converge the parameter estimates had unacceptable levels of bias. In fact, for Models II and III, simply ignoring the sampling structure of the data and using the Baseline Model sometimes provides slightly better parameter estimates, without the convergence problems. Even if the parameter estimates were accurate, sample designs for large, nationally representative data sets are often too complex to model through the use of covariates.

The weighted model (Model I) performs the best out of all of the models tested, but still results in parameter estimates with unacceptably high percentages of bias. The estimated parameters were not that much more accurate than the unadjusted Baseline Model. In fact, Stapleton (2006) found similar results using effective weights in a more general SEM framework and subsequently advised against their use in SEM. It is important to note that, although Mplus was provided with effective weights in Models I and III, the program rescales all weights to relative weights, which sum to the sample size. For a detailed explanation of this, the reader is directed to Asparouhov (2005).

Although it is possible that this rescaling of the weights may have impacted the accuracy of parameter estimates in Models I and III, it seems unlikely, as Model II did not use weights, and also exhibited considerable bias. It is more likely that the mean separation and variance of the manifest variables are to blame. If the distributions of the manifest variables from one class overlap too much with the variables in the second class, the aggregate distribution becomes unimodal, making it difficult for the algorithm to distinguish among the latent classes and thus affecting the accuracy of parameter estimates. That is, if individuals are assigned to the wrong latent class, then class specific parameters will not be estimated accurately. Figures 6 and 7 below provide the histograms of one manifest variable used in this research. Figure 6 illustrates the distribution of this variable, conditional on class, while Figure 7 shows that when these data are combined the resulting distribution looks disturbingly normal.

Figure 6. Distribution of class 1 (top) and class 2 (bottom) manifest variable



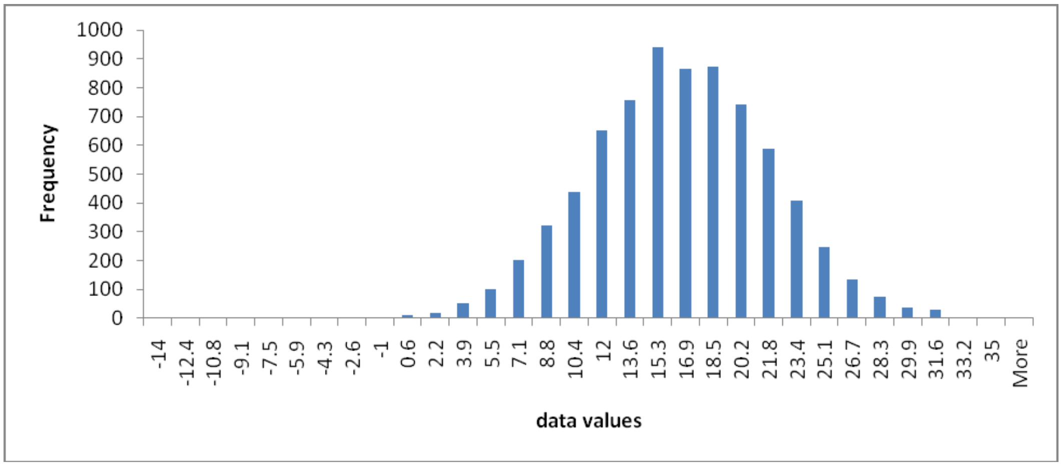
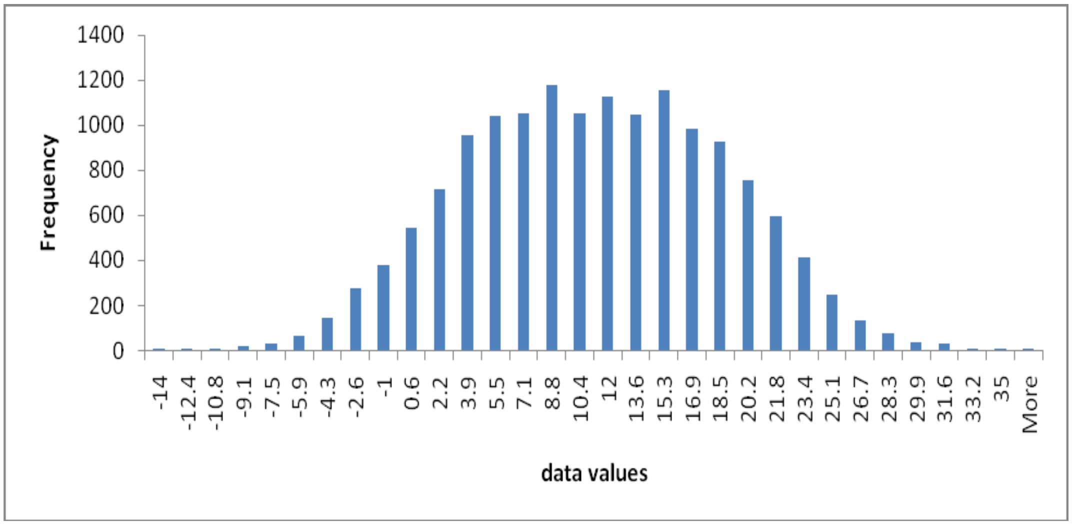


Figure 7. Combined distribution of manifest variable



One way to minimize the problem of overlapping distributions is to ensure that there is more separation between the means. However, the degree of variability around the means also play a factor. If this variability is reduced, less of the distributions would overlap. Unfortunately, neither of these two factors are likely to be within the control of the applied researcher.

A potential way for empirical researchers to improve class assignment is to add class-predicting covariates to the model (Lubke & Muthén, 2007). However, these predictors would make the model more complex, and so convergence problems may result. Although the results reported here only include a handful of approaches to modeling sample data with disproportionate probabilities of selection, the results from this research are not encouraging. In sum, the research reported here seem to indicate that GMM should not be used when data are sampled with disproportionate probabilities. Researchers should therefore attend to the study design and data collection strategies when considering the use of a Growth Mixture Model in the analysis phase. Especially in secondary analyses of nationally representative data sets, it is important to carefully review the documentation pertaining to data collection, especially the use of stratification.

If GMM are used with data sampled with differing probabilities of selection, there are two potential ‘work around’ solutions. One would be to analyze each stratum in a separate Growth Mixture Model, and then weight and combine these separate parameter estimates. The second would be to randomly delete cases from the larger strata until the sample resembles one collected via SRS. These data could then be analyzed with a conventional Growth Mixture Model.

REFERENCES

Allua, S., Stapleton, L. M., & Beretvas, S. N. (2008). Testing latent mean differences between observed and unobserved groups using multilevel factor mixture models. *Educational and Psychological Measurement, 68*, 357-378.

Arminger, G., & Stein, P. (1997). Finite mixtures of covariance structure models with regressors. *Sociological Methods and Research, 26*, 148-182.

Arminger, G., Stein, P., & Wittenburg, M. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrica*, *64*, 475-494.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 12*, 411-434.

Bamber, D., & Santer, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology, 44*, 20-40.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Griffin.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory cases. *Psychological Methods, 8*, 338-363.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3-29.

Bentler, P., & Chou, C.-P. (1987). Practical issues in structural equation modeling. *Sociological Methods & Research, 16*, 78-117.

Boscardin, C. (2001). Strategies for early identification of students with reading difficulties using growth mixture models. (Doctoral dissertation, University of California, Los Angeles, 2001). *Dissertation Abstracts International,* Section A: Humanities and Social Sciences (UMI No. 3026279).

Brogan, D. J. (1998). Pitfalls of using standard statistical software packages for sample survey data. In P. Armitage and T. Colton (Eds.) *Encyclopedia of biostatistics*. West Sussex, England: Wiley.

DeFraine, B., VanDamme, J., & Onghena, P. (2004). *Predicting longitudinal trajectories of adolescent academic self-concept: An application of growth mixture models* (Technical Report 0474). Louvain-la-Neuve, Belgium: Interuniversity Attraction Pole, Belgium Science Policy.

Deming, W. E., & Stephan, F. F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association, 36,* 45–49.

DuMouchel W. H., & Duncan, G. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association, 78*, 535-543.

Duncan, T.E., Duncan, S. C., Strycker, L. A., Okut, H., & Li, F. (2002). Growth mixture modeling of adolescent alcohol use data. Chapter addendum to *An Introduction to latent variable growth curve modeling: Concepts Issues, and Applications*. Eugene, OR: Oregon Research Institute.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

Gagné, P. E. (2004). General confirmatory factor mixture models: A tool for assessing factorial invariance across unspecified populations (Doctoral dissertation, University of Maryland, 2004). *Dissertation Abstracts International, 65,* 1389B.

Gagné, P. E., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, 41, 65-83.

Galton, F. R. S. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London, 45*, 135-145.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24,* 229-252.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identified and unidentified models. *Biometrika, 61*, 215-231.

Greenbaum, P. E., Del Boca, F. K., Darkes, J., Wang, C., & Goldman, M. S. (2005). Variation in the drinking trajectories of freshman college students. *Journal of Consulting and Clinical Psychology, 73*, 229-238.

Hahs, D. (2003, April). *Utilization of sample weights in single level structural equation modeling*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association, 78*, 776-793.

Heeringa, S. D., & Liu, J. (1997). Complex sample design effects and inference for mental health survey data. *International Journal of Methods in Psychiatric Research, 7*, 221-230.

Holt, D., Smith, T. M. F., & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, 143*, 474-487.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329-367.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47,* 663-685.

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

Jeffries, N. O. (2003). A note on 'Testing the number of components in a normal mixture.' *Biometrika, 90*, 991-994.

Kalton, G. (1983). *Introduction to survey sampling.* Thousand Oaks, CA: Sage.

Kaplan, D. (2002). Methodological advances in the analysis of individual growth with relevance to educational policy. *Peabody Journal of Education, 77*, 189-215.

Kaplan, D., & Ferguson, A. (1999). On the utilization of sampling weights in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 305-321.

Kish, L. (1995). *Survey sampling*. New York: Wiley.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, 36*, 1-37.

Korn, E., & Graubard, B. (1995a). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, 158*, 263-295.

Korn, E., & Graubard, B. (1995b). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49*, 291-295.

Kreisman, M. (2003). Evaluating academic outcomes of Head Start: An application of general growth mixture modeling. *Early Childhood Research Quarterly, 18*, 238-254.

Lawly, D. N. & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London: Butterworths.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lee, E. S., Forthofer, R., & Lorimor, R. (1989). *Analyzing complex survey data.* Thousand Oaks, CA: Sage.

Lehtonen, R., & Pahkinen, E. J. (1994). *Practical methods for design and analysis of complex surveys.* New York: Wiley.

Li, F., Barerra, M., Hops, H., & Fisher, K. J. (2002). The longitudinal influence of peers in the development of alcohol use in late adolescence: A growth mixture analysis. *Journal of Behavior Medicine, 25*, 293-315.

Li, F., Duncan, T. E., Duncan, S. C., & Acock, A. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 493-530.

Little, R. J. A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Society, 78*, 596-603.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.

Lohr, S. L. (1999). *Sampling: Design and analysis.* New York: Duxbury Press.

Longford, N. (1995). *Model-based methods for analysis of data from the 1990 NAEP trial state assessment.* Princeton, NJ: Educational Testing Service.

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

Lubke, G., Muthén, B., & Larsen, K. (2002). *Empirical identifiability of factor mixture models.* Unpublished manuscript.

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

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110-113.

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

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

Muthén, B. (1991). Analysis of longitudinal data using latent variable models with varying parameters. In L. Collins & J. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions,* (pp. 1-17). Washington, DC: American Psychological Association.

Muthén, B. (1992). *Advances in multi-level psychometric models: Latent variable modeling of growth with missing data and multilevel data.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Muthén, B. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research, 22*, 376-398.

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

Muthén, B. (2001b). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class, latent growth modeling. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change,* (pp. 291-322). Washington, DC: American Psychological Association.

Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81-117.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling. *Psychological Methods, 8*, 369-377.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Thousand Oaks, CA: SAGE Publications.

Muthén, B., Brown, C. H., Khoo, S., Yang, C. C., & Jo, B. (1998). *General growth mixtures modeling of latent trajectory classes: Perspectives and prospects*. Paper presented at the meeting of the Prevention Science and Methodology Group, Tempe, AZ.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C. et al. (2002). General growth mixture modeling for randomized preventative interventions. *Biostatistics*, *3*, 459-475.

Muthén, B., Khoo, S., & Francis, D. (1998). *Multi-stage analysis of sequential developmental processes to study reading progress: New methodological* *developments using general growth mixture modeling.* University of California, Los Angeles: Center for the Study of Education.

Muthén, B., Khoo, S., Francis, D., & Boscordin, K. (2003). Analysis of reading skills development from kindergarten through first grade: An application of growth mixture modeling to sequential process. In S. R. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues and applications,* (pp. 71-89). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Muthén, B., & Muthén, L. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 4,* 599-620.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological Methodology,* (pp. 216-316). Washington, DC: American Sociological Association.

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

Muthén, L., & Muthén, B. (2005). *Mplus user's guide: Third edition*. Los Angeles, CA: Muthén & Muthén.

Nagin, D. (1999). Analyzing developmental trajectories: A semi parametric, group-based approach. *Psychological Methods, 4*, 139-157.

Nagin, D., & Tremblay, R. (1999). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Development, 70,* 1181-1196.

Nathan, G., & Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Association, 42*, 377-386.

Neuman, R. J., Todd, R. D., Heath, A. C., Reich, W., Hudziak, J. J., Buchalz, K. K. et al. (1999). The evaluation of ADHD typology in three contrasting samples: A latent class approach. *Journal of the American Academy of Child Adolescent Psychiatry, 38,* 25-33.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*, 558-606.

Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 535-569.

Patterson, B. (1998). Latent class analysis of sample survey data. (Doctoral Dissertation, University of Maryland, College Park, 1998). *Dissertation*

Abstracts International, Section A: Humanities and Social Sciences (UMI No. 9920986).

Patterson, B., Dayton, C. M., & Graubard, B. I. (2002). Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association, 97*, 721-729.

Paxton, P., Curran, P., Bollen, K., Kirby, J., & Chem., F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 287-312.

Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*, 317-337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, 60*, 23-56.

Potthoff, R., Woodbury, M. A., & Manton, K. G. (1992). Equivalent sample size and equivalent degrees of freedom: Refinements for inference using survey weights under subpopulation models. *Journal of the American Statistical Association, 87*, 383-396.

Rasmussen, E. R., Neuman, R. J., Heath, A. C., Levy, F., Hay, D. A., & Todd, R. D. (2002). Replication of the latent class structure of Attention Deficit/Hyperactivity Disorder (ADHD) subtype in a sample of Australian twins. *Journal of Child Psychology and Psychiatry, 43*, 1018-1028.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In A. Sayer & L. Collins (Eds.), *New methods for the analysis of change.* Washington, DC: American Psychological Association.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1,* 381-397.

SAS Institute (2004). *SAS procedures guide*. Cary, NC: SAS Institute.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys*. New York: Wiley.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

SPSS (1999). SPSS Applications Guide 10.0. SPSS Inc., Chicago.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 475-502.

Stapleton L. (2006a) An assessment of practical solutions for Structural Equation Modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal, 13*, 28-58.

Stapleton, L. (2006b). Using multilevel Structural Equation Modeling techniques with complex sample data. In G. Hancock & R. Mueller (Eds.), *Structural Equation Modeling: A Second Course*. Charlotte, NC: IAP

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Sturgis, P. (2004). Analyzing complex survey data: Clustering, stratification and weights. *Social Research Update, 43*, 1-6.

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*, 495-506.

Thurstone, L. L. (1935). *The vectors of the mind*. Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: University of Chicago Press.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in Growth Mixture Models. In G. Hancock & K. Samuelson (Eds.), *Advances in latent variable mixture models*. Charlotte, NC: IAP.

U.S. Department of Education (2003). *NCES handbook of survey methods: Technical report*. Washington, DC: National Center of Education Statistics.

Van Lier, P., Muthén, B., van der Sar, R. M., & Crijnen, A. (2004). Preventing disruptive behavior in elementary schoolchildren: Impact of a universal classroom-based intervention. *Journal of Consulting and Clinical Psychology, 72*, 467-478.

Wang, M. & Bodner, T. E. (2007). Growth Mixture Modeling: Identifying and Predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, *10*, 635-656.

Wedel, M., Hofstede, F., & Steenkamp, J.-B. (1997). *Mixture model analysis of complex samples*. University of Groningen: Research School Systems Organization and Management.

Wiesner, M., & Capaldi, D. (2003). Relations of childhood and adolescent factors to offending trajectories of young men. *Journal of Research in Crime and Delinquency, 40*, 231-262.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15,* 315-422.

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

Yuan, K., & Bentler, P. (2006). Asymptotic robustness of standard errors in multilevel structural equation models. *Journal of Multivariate Analysis, 97*, 1121-1141.

Zheng, J., & Frey, C. (2004). Quantification of variability and uncertainty using mixture distributions: Evaluation of sample size, mixing weights, and separation between components. *Risk Analysis, 24*, 553-566.