Graduate Theses and Dissertations   Graduate School

6-9-2003

Estimates of Statistical Power and Accuracy for Latent Trajectory Class Enumeration in the Growth Mixture Model

Eric C. Brown
University of South Florida

Follow this and additional works at: https://scholarcommons.usf.edu/etd

Part of the American Studies Commons

Estimates of Statistical Power and Accuracy for Latent Trajectory Class

Enumeration in the Growth Mixture Model

by

Eric C. Brown

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Educational Measurement and Research
College of Education
University of South Florida

Co-Major Professor: Robert F. Dedrick, Ph.D.
Co-Major Professor: Paul E. Greenbaum, Ph.D.

C. Hendricks Brown, Ph.D.
John Ferron, Ph.D.
Cynthia Parshall, Ph.D.

Date of Approval:
June 9, 2003

Keywords: Monte Carlo simulation, structural equation model, noncentral chi-square distribution, longitudinal design, sample size determination

Dedication

To my parents, Cecilia Jones and Richard W. Brown, with love and thanks.

Acknowledgments

My sincerest appreciation goes to those who made this dissertation

possible: Prof. Mark Amen, Director of the Globalization Research Center at the

University of South Florida; Prof. Robert Friedman, former colleagues, and

friends at the Department of Child and Family Studies, Louis de la Parte Florida

Mental Health Institute (University of South Florida); Dr. Wayne Holden of ORC

Macro International; Prof. Bengt Muthén, Dr. Chen-Pin Wang, Katherine Masyn,

and members of the Prevention Science Methodology Group; and Dr. M. Lee Van

Horn and Charles B. Fleming at the Social Development Research Group

(University of Washington). Also, my deepest heartfelt thanks to Mabel for her

help and support throughout.

Table of Contents

Table of Contents

List of Tables

List of Figures

iv

Estimates of Statistical Power and Accuracy for Latent Trajectory
Class Enumeration in the Growth Mixture Model

Eric C. Brown

ABSTRACT

This study employed Monte Carlo simulation to investigate the ability of the growth

mixture model (GMM) to correctly identify models based on a "true" two-class pseudo-

population from alternative models consisting of "false" one- and three-latent trajectory

classes. This ability was assessed in terms of statistical *power*, defined as the proportion

of replications that correctly identified the two-class model as having optimal fit to the

data compared to the one-class model, and *accuracy*, which was defined as the proportion

of replications that correctly identified the two-class model over both one- and three-class

models. Estimates of power and accuracy were adjusted by empirically derived critical

values to reflect nominal Type I error rates of $\alpha = .05$. Six experimental conditions were

examined: (a) standardized between-class differences in growth parameters, (b)

percentage of total variance explained by growth parameters, (c) correlation between

intercepts and slopes, (d) sample size, (e) number of repeated measures, and (f) planned

missingness. Estimates of statistical power and accuracy were related to a measure of the

degree of separation and distinction between latent trajectory classes (λ^2), which

approximated a chi-square based noncentrality parameter. Model selection relied on four

criteria: (a) the Bayesian information criterion (BIC), (b) the sample-size adjusted BIC

(ABIC), (c) the Akaike information criterion (AIC), and (d) the likelihood ratio test

(LRT). Results showed that power and accuracy of the GMM to correctly enumerate latent trajectory classes were positively related to greater between-class separation, greater proportion of total variance explained by growth parameters, larger sample sizes, greater numbers of repeated measures, and larger negative correlations between intercepts and slopes; and inversely related to greater proportions of missing data. Results of the Monte Carlo simulations were field tested using specific design and population characteristics from an evaluation of a longitudinal demonstration project. This test compared estimates of power and accuracy generated via Monte Carlo simulation to estimates predicted from a regression of derived λ^2 values. Results of this motivating example indicated that knowledge of λ^2 can be useful in the two-class case for predicting power and accuracy without extensive Monte Carlo simulations.

Chapter One

Introduction

Researchers have long understood the importance of analyzing change in

individuals over time. Recently, methods to analyze change in individuals have evolved

into sophisticated latent variable techniques that incorporate change in individuals into a

broader context of change within populations. One such technique is latent growth

modeling (LGM), which relies on individual growth trajectories to determine patterns of

change in larger groups of individuals. A primary assumption of LGM is that the growth

trajectory that represents change in the dependent indicator variables is modeled as a

single population distribution and that any nonrandom deviation from the underlying

population distribution must be modeled explicitly by covariates included in the study

design (e.g., intervention status, age, gender). However, the presence of multiple

distributions in a population can result in unobserved heterogeneity in the distribution of

individual growth trajectories around a single population distribution and, consequently,

requires a *mixture modeling* approach. In response to this, researchers (e.g., Muthén &

Shedden, 1999; Nagin, 1999) have recently developed models, such as the general

growth mixture model and the semiparametric group-based model, that are based on the

assumption that individuals' growth trajectories can be simultaneously modeled by more

than a single population distribution. In these models, the degree to which an individual's

growth trajectory can be modeled by a particular population distribution is considered to

be stochastic and the probability of being correctly modeled by a particular distribution is quantified explicitly. This gives rise to the identification of *latent trajectory classes* of individuals, which constitute likely groupings of individuals with similar patterns of change.

The utility of these models lies in their ability to identify typologies for latent trajectory classes. Consequently, these typologies can help researchers develop interventions tailored to the unique needs of these groups. If the utility of these models is to be maximized, then the reliable identification (i.e., enumeration) of latent trajectory classes must first be established. This is complicated, however, by the limited distributional theory underlying statistical tests involving models with competing numbers of latent trajectory classes. Furthermore, existing model selection criteria used to determine the optimal number of latent trajectory classes in longitudinal data have not been examined fully with regard to varying population distributions or design constraints. Therefore, the purpose of this study was to use Monte Carlo simulation of sample data drawn from a population containing a mixture of two distinct distributions to ascertain the conditions necessary to insure adequate statistical power and accuracy to detect multiple latent trajectory classes. An additional purpose of this study was to evaluate the various criteria by which models are identified in terms of their Type I error rates. A final purpose of this study was to demonstrate, through application of a motivating example, the process by which researchers can derive estimates of power and accuracy for enumeration of latent trajectory classes for specific population and design characteristics.

The Latent Growth Model

One popular technique to model individual change over time is the latent growth model (Duncan, Duncan, Strycker, Li, & Alpert, 1999; Willett & Sayer, 1994). In the simplest form of the LGM, change over time is modeled as a linear function of individual growth trajectories, denoted as

$$y_{it} = \eta_{0i} + \eta_{1i}(x_t) + \varepsilon_{it},$$

where values for the dependent indicator variable (y_{it}) are modeled as a linear function of an individual's specific intercept (η_{0i}), linear slope (η_{1i}), and a random error term ε_{it}, $\sim N(0, \sigma^2)$, where σ^2 represents the residual variance. Variable x represents repeated measurement occasions where, for example, $t = 0, 1, 2, \ldots, T$. Within the broader framework of the structural equation model (SEM), this is referred to as the measurement model (Muthén, 2001b). In the framework of the multilevel or hierarchical linear model (HLM), this is referred to as the Level-1 model (Goldstein, 1995; Raudenbush & Bryk, 2002).

In turn, individuals' specific intercept and slope parameters can be modeled as a linear function of latent growth factors and covariates, represented by

$$\eta_{0i} = \alpha_0 + \gamma_0 w_i + \zeta_{0i}$$

and

$$\eta_{1i} = \alpha_1 + \gamma_1 w_i + \zeta_{1i},$$

where α_0 and α_1 represent population-average intercept and slope parameters, respectively, w represents a time-invariant covariate (e.g., intervention status, age, gender), and γ_0 and γ_1 represent the effects of w on an individual's specific intercept and slope parameters. Residual error terms, ζ_0 and ζ_1, are bivariate normally distributed, $\zeta_{0i} \sim$

N (0, σ_0^2) and $\zeta_{1i} \sim$ N (0, σ_1^2), where σ_0^2 and σ_1^2 represent residual variances for each

growth parameter, respectively. This is referred to as the structural model in SEM and the

Level-2 model in the multilevel model or HLM. Muthén and Khoo (1998, p. 82) provide

a diagrammatic example of the LGM, reproduced in Figure 1. In this figure, squares

represent the observed variables and circles represent latent variables. Additionally, *T*

repeated measures of indicator variable y_t are modeled as a function of correlated latent

growth factors η_0 and η_1 (with loadings for each growth factor shown in the diagram) and

time-invariant covariate *w*.

The Latent Variable Mixture Model

As mentioned earlier, population-average intercept and slope parameters in the

LGM (α_0 and α_1, respectively) are based on the assumption that the dependent indicator

variable (y_{it}) is represented by a single population distribution, characterized by a specific

underlying moment-generating function (e.g., mean, variance, skewness, and kurtosis).

When a population contains more than one underlying distribution for the y_{it} variable,

mixture distributions are said to exist. The latent variable mixture model (LVMM; Everitt

& Hand, 1981; McLachlan & Basford, 1988; McLachlan & Peel, 2001; Muthén &

Shedden, 1999; Titterington, Smith, & Makov, 1985) takes mixture distributions into

account by partitioning the unobserved heterogeneity in data (i.e., heterogeneity that is

not readily captured by covariates) and assigning probabilities of membership to latent

classes for each individual, with the overall number of classes in the model being

specified by the researcher using a particular model selection criterion and decision rule.

These *posterior probabilities* of latent class membership typically vary across latent

classes (summing to unity across all latent classes within a model) and, therefore, can

have differential relationships with indicator variables and covariates, as well as moderate the relationship between indicators and covariates. Muthén (2001a, p. 34) provides a diagrammatic example of these relationships, reproduced here in Figure 2. As shown in the figure, multiple indicator variables (y_1, y_2, y_3, … y_n) can be regressed on a latent class variable (c) in a manner analogous to the traditional factor analytic model. When indicator variables are continuous, parameter estimates represent the mean for each indicator variable and the modeling is referred to as a latent profile analysis. When indicator variables are binary, parameter estimates represent the probability for each indicator variable and the modeling is referred to as a latent class analysis (Muthén, 2001a).

The General Growth Mixture Model

Despite their utility, LVMMs are insufficient to capture unobserved heterogeneity that is best represented as a dynamic process (e.g., during a developmental period of an individual's life). To remedy this, the LVMM can be combined with the LGM in a generalized modeling framework referred to as the general growth mixture model (GGMM; Muthén, 2001b; Muthén et al., 2002). In the GGMM, latent trajectory classes are identified by assigning posterior probabilities of latent class membership based on individuals' growth trajectories. Moreover, the GGMM represents a family of latent variable models (e.g., factor analytic, latent class, latent transition, and growth mixture models; see Muthén & Curran, 1997; Muthén & Muthén, 2000) and, consequently, has been used in a variety of applications, such as studies of reading development (Dedrick, Greenbaum, & Vaughn 2001; Muthén, Khoo, Francis, & Boscardin, 2000); school expulsion (Petras et al., 2001); emotional/behavioral problems (Greenbaum, Dedrick, &

Brown, 2000a, 2000b); alcohol use (Colder, Campbell, Ruel, Richardson, & Flay, 2002;

Greenbaum, Del Boca, Darkes, & Goldman, 2001; Li, Duncan, & Hops, 2001; Oxford et

al., 2003; Tucker, Orlando, & Ellickson, 2003); adolescent smoking (Colder et al., 2001);

and service delivery interventions (Greenbaum & Brown, 2002). Currently, the GGMM

can be implemented using the Mplus computer program (Muthén & Muthén, 2001).

A diagrammatic example of the GGMM is reproduced in Figure 3 (Muthén &

Muthén, 2001, p. 210; factor loadings and error terms omitted for simplicity). This figure

represents the relationships among the repeatedly measured indicator variable (y_t),

intercept and slope growth factors (η_0 and η_1), latent class variable (c), time-invariant

covariate (w), and categorical outcome variable (u). The model allows for within- and

between-class variability while examining unobserved heterogeneity in growth

trajectories. Additionally, the model can examine time-variant and -invariant covariates

(e.g., treatment and intervention effects), mediators, moderators, and proximal or distal

outcomes, all under various longitudinal structures (e.g., linear, quadratic, nonequidistant,

autocorrelated) and designs (e.g., cohort-sequential, planned missingness). When

categorical outcome variables are omitted, the GGMM reduces to the growth mixture

model (GMM), as reproduced in Figure 4 (Muthén & Muthén, 2001, p. 207).

Model Selection

Despite its utility, several issues still remain to be examined in the GGMM.

Primary among these is the issue of model selection. Methods to assess model fit in

traditional latent variable models typically rely on the asymptotic properties of a

particular parametric distribution (e.g., chi-square) comparing fitted models to either a

fully-restricted comparison model (i.e., all correlations constrained to zero) or a fully-

unrestricted model (i.e., relying on the population variance-covariance matrix). From this,

a variety of goodness-of-fit indices have been developed to provide benchmarks for

reliable model selection, including, for example, the Goodness-of-Fit Index (GFI),

Normed Fit Index (NFI), Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and

Root Mean Square Error of Approximation (RMSEA). Comparison of nested latent

variable models is routinely conducted using the likelihood ratio test (LRT) where a more

complex model (H_1) is tested for significant improvement in fit compared to a reduced

model (H_0). The LRT is described in Muthèn and Muthèn (2001, p. 283) as $2nF_{ML}(\pi)$

where π denotes the maximum likelihood estimate under H_0. The function $F_{ML}(\pi)$ is

expressed as

$$F_{ML}(\pi) = - logL_{H0} / n + logL_{H1} / n.$$

From this function, –2 times the difference in log likelihoods between H_0 and H_1 (also

referred to as the deviance) can be tested using a chi-square distribution with degrees of

freedom equal to the difference in the number of freely estimated parameters between

models.

However, in the case of the GGMM, the distributional properties of likelihood

functions, as they relate to model fit, have not yet been rigorously defined. Use of the

LRT to compare growth mixture models under the assumption of a true null hypothesis of

a single population distribution (i.e., one latent trajectory class) is problematic "given that

this involves inadmissible parameter values of zero class probabilities" (Muthèn &

Muthèn, 2001, p. 292). That is, "the true value of the parameter vector under H_0 [lies] on

the boundary of the parameter space…" and, therefore, regularity conditions necessary

for the appropriate estimation of the likelihood function are not satisfied (McLachlan &

Peel, 2001, p. 185; also see Clogg, 1995, for theory underlying the inappropriateness of the LRT in mixture distributions). As the distributional theory underlying statistical comparison of mixture models with different numbers of latent classes has yet to be defined, traditional fit indices such as the GFI, NFI, TLI, CFI, and RMSEA cannot be used for model selection.

Alternative methods to determine the number of latent classes in mixture distributions models include examination of residual diagnostics (Lindsay & Roeder, 1992; Wang, Brown, Carlin, & Dagne, 2001), the method of moments (Dacunha-Castelle & Gassiat, 1997; Furman & Lindsay, 1994; Heckman, Robb, & Walker, 1990), bootstrapping techniques (McLachlan, 1987; McLachlan & Peel, 1997), cross-validation (Zucchini, 2000), and discrepancy risk model selection (Golden, 2000). Despite the variety of methods, most applied research has relied on penalized information criteria for latent class enumeration, which adds a penalty to the estimated log likelihood corresponding to the number of estimated parameters in the model. Although a wide variety of such criteria exist, the most frequently used are the Akaike information criterion (AIC; Akaike, 1973), Bayesian information criterion (BIC; Schwarz, 1978), and the sample-size adjusted BIC (ABIC; Sclove, 1987).

As a simplification of the Kullback-Leibler (1951) discrepancy between a true population distribution and its modeled approximation, the AIC is defined as

$$AIC = -2\ logL + 2\ r,$$

where r indicates the number of free parameters in the model. The BIC is similar to the AIC, however, it originates from a Bayesian perspective and theoretically provides more

asymptotically consistent estimates by incorporating the natural log of sample size (n) in

the penalty:

$$BIC = -2 \; logL + r \; ln \; n.$$

In 1987, Sclove considered another adaptation to the penalty term (originally introduced

by Rissanen, 1978) to further improve the consistency of the criterion in finite samples,

replacing n with

$$n^* = (n + 2) / 24.$$

This is referred to as the sample-size adjusted BIC (ABIC). Smaller values for the AIC,

BIC, and ABIC indicate better fitting models. Despite their applicability, these

information criteria are limited by their inability to produce inferential probability

statements, in other words, of a "number that quantifies the confidence in the result, such

as a p value" (McLachlan & Peel, 2001, p. 184). Additionally, these criteria also have

been limited by the general lack of knowledge regarding their behavior under various

parameterizations (e.g., sample size).

Statistical Power

      Confidence in the use of these criteria can be gained through power analysis

(Cohen, 1977). Generally, *statistical power* can be thought of as "…the probability of

producing reliable findings…if a researcher's hypothesis is true" (Dennis, Lennox, &

Foss, 1997, p. 367). More specifically, power can be defined as the ability to correctly

reject a false null hypothesis, or in other words, one minus the probability of making a

Type II error (i.e., power = 1 - β).

      With regard to latent variable models, methods to estimate statistical power have

relied on both the asymptotic properties of existing parametric distributions (see

MacCallum, Browne, & Sugawara, 1996; Muthén & Curran, 1997; Saris & Satorra, 1993; Satorra & Saris, 1985) and Monte Carlo simulation (see Gerbing & Anderson, 1992; Ichikawa & Konishi, 1999; Olmos & Hutchinson, 1998; Xitao, Thompson, & Wang, 1999; Xitao & Wang, 1998). In terms of model selection, both methods relate to statistical power as the ability to reject a false null hypothesis of a misspecified model in favor of a true alternative hypothesis of a correctly specified model. However, in the former method, statistical power is assessed by the difference between nested models, expressed as chi-square units or some other discrepancy function such as the RMSEA. This difference can be represented as a noncentrality parameter based on the noncentral chi-square distribution. In turn, power estimates to reject the erroneous null hypothesis can be derived from tables of this distribution (e.g., Haynam, Govindarajulu, & Leone, 1973).

Alternatively, the Monte Carlo simulation approach to estimating statistical power provides estimates of statistical power that are derived from empirically derived distributions based on repeated draws (i.e., replications) of simulated data generated from particular experimental conditions of interest (Mooney, 1997). Within the context of model selection, power estimates are based on the proportion of replications that indicate acceptable fit, with greater numbers of replications resulting in smaller confidence intervals around point estimates. When particular parameterizations cannot be covered by existing asymptotic theory, or when existing theory has not yet been fully developed--as in the case of mixture distributions--Monte Carlo simulation can be used to determine the performance of model selection criteria.

Statement of the Problem

As noted earlier, the GMM provides a method to identify latent trajectory classes of individuals not explicitly identified by covariates included in longitudinal designs. As unexplained heterogeneity in data increases the probability of Type II errors (e.g., failing to detect a statistically significant covariate effect in sample data), the power to accurately identify the correct number of latent trajectory classes first must be established. However, the performance of existing model selection criteria to correctly enumerate latent trajectory classes in the GMM is only just beginning to be examined. For example, Masyn (2001) and Masyn and Brown (2001) utilized Monte Carlo simulation (number of replications = 200 to 500) to assess the impact of various longitudinal mixture distributions and design conditions on the performance of several model selection criteria. Although these studies examined a relatively restricted range of experimental conditions (e.g., sample size, repeated measurement occasions, differences in class trajectories), results suggested that the conditions necessary to insure adequate statistical power can be expensive in terms of the amount of overall data that are needed. As longitudinal studies are costly and difficult to implement, it becomes particularly important to gain detailed knowledge of the specific research conditions necessary to develop appropriate and efficient study designs.

Purpose of the Study

In practice, the process of identifying multiple latent trajectory classes is sequential. That is, the applied researcher first attempts to reject the assumption of a single population distribution underlying the data by testing a one-class model against an alternative two-class model, with the two-class model implicitly assuming a mixture of

two latent population distributions. The purpose of this study, therefore, was to evaluate

the ability of the GMM to recover specific latent trajectory classes from simulated data

with a known mixture of two latent trajectory classes. This ability was expressed in terms

of statistical power, operationalized as the proportion of the 500 replications in each

experimental condition in which the two-class model was correctly identified as having

better fit than the competing one-class model. An additional purpose of this study was to

examine the *accuracy* of latent trajectory class enumeration in the GMM, with accuracy

defined as the proportion of the 500 replications in each experimental condition in which

the two-class model was correctly identified as having the best fit compared to both one-

and three-class models. As no single criterion has been clearly identified as superior,

several model selection criteria (i.e., AIC, BIC, ABIC, and LRT) were used. Although

minimum values for the AIC, BIC, and ABIC are typically used in practice to indicate

best fit, for the current study, critical values for these criteria were identified at an

effective .05 Type I error rate based on empirical distributions resulting from Monte

Carlo simulations of "true" one-class population distribution versus "false" two-class

models. Additionally, it should be noted that determination of statistical power was

conducted under a true alternative hypothesis of two latent trajectory classes and

consequently avoided violating regularity conditions necessary for the application of the

LRT.

      Six distinct experimental conditions, related to differences in mixture distributions

in the population and longitudinal design characteristics, were examined. Specifically,

this study provided estimates of statistical power and accuracy as functions of: (a)

standardized differences in intercept and slope parameters between latent trajectory

classes, (b) percentage of variance explained by intercept and slope (growth) parameters, (c) correlation between intercept and slope parameters, (d) sample size, (e) number of repeated measures, and (f) proportional planned missingness. Conditions (a), (b), (d), and (e) were theorized to be positively related to statistical power and accuracy; that is, greater between-class separation in intercepts and slopes, greater percentages of explained variance, increased sample sizes, and increased numbers of repeated measures were predicted to result in improved detection of latent trajectory classes. Alternatively, conditions (c) and (f) were theorized to be inversely related to power and accuracy; that is, larger positive correlations between intercept and slope parameters (within classes) and greater percentages of missing data were predicted to result in poorer detection of latent trajectory classes.

Changes within these experimental conditions impacted power and accuracy through the likelihood functions generated by the GMM in parameter estimation. For example, increased separation in intercepts and slopes, expressed as either larger between-class differences in growth parameter means or as larger negative within-class correlations between growth parameters, should result in a greater likelihood function for the "true" two-class model relative to the erroneous one- and three-class models. Greater values for the two-class log likelihood should translate into smaller AIC, BIC, and ABIC values and, consequently, greater estimates of power and accuracy. For the LRT, increased separation in growth parameters similarly should result in greater differences in log likelihoods between competing models with the end result again being greater power and accuracy. Increasing the percentage of variance explained by growth parameters is equivalent to reducing the degree of measurement error in the model, which similarly

should result in a greater improvement in model fit. Examined design conditions are thought to be related to the amount of information that is available for parameter estimation either through increasing the number of participants or time points, or decreasing the overall percentage of missing data—all resulting in parameter estimates that are closer to true population values and, consequently, improved model fit. Simply put, the power and accuracy of the GMM to enumerate latent trajectory classes is theorized to be positively related to the amount of information available for parameter estimation and the precision of that information.

The final goal of this study was to provide a procedure by which researchers can produce estimates of power and accuracy for latent trajectory class enumeration for various populations and designs outside the scope of the study. Specifically, the Monte Carlo-derived estimates of power and accuracy, based on the two-class population, were used to develop a general functional relationship to an overall measure of the degree of separation between the latent trajectory classes. This measure was defined as the average of –2 times the difference in log likelihoods between one-and two-class models across n replications, minus degrees of freedom equal to the difference in the number of freely estimated parameters, and was modeled as an approximation to a noncentral chi-square parameter (represented as λ^2 in this study). As a test of this relationship, estimates of power and accuracy derived from this general function were compared to empirically generated estimates of power and accuracy resulting from a case-specific Monte Carlo simulation, which used the particular population and design characteristics of a motivating example.

A Motivating Example

In addition to Monte Carlo simulation, actual data from an evaluation of the Comprehensive Community Mental Health Services for Children and Their Families Program (conducted by ORC Macro International) was used as a motivating example. The purpose of the evaluation was to test the effectiveness of a systemic intervention consisting of an integrated *system-of-care* (Stroul & Friedman, 1986) approach to the delivery of children's mental health services in Stark County, Ohio versus a traditional service delivery approach in Mahoning County, Ohio. As part of this evaluation, measures of delinquent behavior from 442 children and adolescents were obtained using the Delinquent Behavior syndrome scale of the Child Behavior Checklist/4-18 (Achenbach, 1991) at five equally spaced time points (i.e., at intake into service and every six months thereafter). Although results of a latent growth model of the entire sample of children were somewhat ambiguous (i.e., no clear intervention effect), an application of the general growth mixture model identified four distinct latent trajectory classes (Greenbaum & Brown, 2002). Moreover, a statistically significant ($p < .05$) intervention effect was found in a class of children whose average delinquency trajectory declined from a clinical to nonclinical level; that is, in the specific subpopulation of children that the intervention was originally designed to target. In this study, data from the system-of-care evaluation were used to illustrate the process of estimating statistical power *post hoc* and to "field test" the results of the Monte Carlo simulations.

Research Questions

Given the current need to expand the knowledge of the GMM's ability to accurately enumerate latent trajectory classes, this study addressed the following research questions:

1. For a particular longitudinal design, how do standardized between-class differences in intercept and slope parameters (i.e., ds = 0.30, 0.50, 0.80, 1.00, and 1.20) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?

2. For a particular longitudinal design, how does the percentage of variance explained by intercept and slope parameters (i.e., 40%, 50%, 60%, 70%, and 80%) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?

3. For a particular longitudinal design, how does the within-class correlation between intercept and slope parameters (rs = -.60, -.30, .00, .30, .60) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?

4. For a particular population of individuals, how does sample size (i.e., Ns = 200, 400, 600, and 800), relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?

5. For a particular population of individuals, how does the number of repeated measures (i.e., 3, 4, 5, 6, 9, and 12 time points) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?

6. For a particular population of individuals, how does planned missing data (i.e., 25%, 50%, 75%, and 100% of cases missing either (a) every other time point or (b) every second and third time point, relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?

7. Which model selection criterion (among the AIC, BIC, ABIC, and LRT) possesses the largest effective Type I error rates, and under which design and population conditions are Type I error rates the largest?

8. How do statistical power and accuracy relate to the overall separation between one- and two-latent trajectory class models, as measured by a noncentrality parameter (λ^2)?

Answers to Questions 1 through 3 provide a basis for researchers to develop designs *a priori* that reliably identify latent trajectory classes in longitudinal data or assess existing longitudinal analyses *post hoc* for power or accuracy. Answers to Questions 4 through 6 relate to latent trajectory class enumeration to population conditions that researchers must ultimately confront in conducting longitudinal studies. Question 7 relates to the choice of model selection criteria to correctly enumerate latent trajectory classes, including the ability to estimate power and accuracy at a common Type I error rate. Finally, Question 8 relates to the development of a procedure by which estimates of statistical power and accuracy can be generalized across multiple experimental conditions and experimental conditions not examined in the current study.

Chapter Two

Literature Review

This chapter provides background on the Monte Carlo simulation methods used in this study, followed by a discussion of literature that used Monte Carlo simulation in classification of growth trajectories or enumeration of latent classes. The chapter concludes with examples of several studies that have experienced disparate results using established model selection criteria in applications of the growth mixture model.

Monte Carlo Simulation

As Mooney (1997) points out in his monograph on Monte Carlo simulation, researchers often are faced with situations in which the distributional properties of a particular statistic (e.g., model selection criterion) have yet to be fully developed. For example, use of the LRT in determining the number of latent classes in data containing mixture distributions continues to be problematic due to the inability of existing asymptotic theory to maintain the regularity conditions necessary for its appropriate application (McLachlan & Peel, 2001). The absence of rigorous theory for these types of distributions requires the use of "brute force" empirical observation to discover their properties, and with the aid of high-speed computer processing, Monte Carlo simulation provides an efficient method to accomplish this.

Essentially, Monte Carlo simulation relies on repeated random sampling of data derived from a population with a known parametric distribution (i.e., a *pseudo-*

population). The observed frequency distribution of the statistic of interest provides a

basis for estimating the density function of that statistic in the population, with the

precision of such estimates being directly related to the number of draws from the

pseudo-population. The procedure for conducting a Monte Carlo simulation study first

requires that the distributional characteristics of the pseudo-population (e.g., mean,

variance-covariance structure, skewness, kurtosis) be specified in terms of a statistical

(i.e., computer) algorithm that models both deterministic (e.g., mean growth trajectories)

and stochastic (e.g., individual deviation from mean growth trajectories) components.

Second, the conditions by which the random data are drawn from the pseudo-population

(e.g., sample size, number of repeated measures, missing data patterns) need to be

decided and built into the sampling algorithm. From this, *t* samples are drawn, where *t*

indicates the number of *trials* or *replications*. The frequency distribution of the examined

statistic across *t* replications then can be mapped out and, consequently, decision rules

can be implemented to determine the performance of the statistic under the specified

conditions. Finally, the characteristics of the pseudo-population and random samples can

be manipulated experimentally to assess the behavior of the statistic across various

conditions.

Monte Carlo Simulation Studies

      Monte Carlo simulation can be used not only to conduct inference where theory is

weak or nonexistent, but also can be used to (a) test a variety of plausible hypotheses, (b)

assess the impact of violations to distributional assumptions, (c) assess the quality of

inferential methods, and (d) compare the properties of various estimators (Mooney,

1997). Examples of recent research in structural equation modeling that have employed

Monte Carlo simulation toward these ends include: Cudeck and Henly (1991); Curran, West, and Finch (1996); Dumenci and Windle (2001); Haughton, Oud, and Jansen (1997); Hu and Bentler (1998); Ichikawa and Konishi (1999); Muthén and Curran (1997); Olmos and Hutchinson (1998); Tanguma (2001); and Xitao, Thompson, and Wang (1999).

For example, Dumenci and Windle (2001) used Monte Carlo simulation to evaluate a cluster analytic method (i.e., Ward's method of squared error criterion) of classifying individual longitudinal growth trajectories. The pseudo-population for this study consisted of various trajectory shapes (i.e., no change, increasing, and decreasing linear and quadratic growth factors), proportions of error variance to observed variance (i.e., .20 and .50), and distances between clusters (1 through 4 standard deviation units between intercept factors). Additional experimental conditions included different mixing proportions (i.e., proportion of cases assigned to each class; πs = .25 and .50) and type of input matrix (e.g., squared Euclidean distance, Mahalanobis distance). All samples were derived from a latent growth model (LGM) consisting of 2,000 cases measured across four time points. The ability of the cluster procedure was assessed using the 90th percentile of the Kappa (κ) distribution (κ = .864) as a benchmark for adequate recovery of simulated growth trajectories. Results of the study averaged across all conditions revealed that overall recovery was "relatively poor" (median κ = .25), requiring a difference of four standard deviations units between intercept means to achieve an average κ = .50. Moreover, recovery of growth trajectories was negligibly related to mixing proportions. The study, however, did not examine enumeration of latent classes (e.g., exploring erroneous one or three cluster solutions as alternatives).

Examples of studies that used Monte Carlo simulation to assess the number of

latent classes in mixture distributions include: Biernacki, Celeux, and Govaert (1999);

Chuang and Mendell (1997); Everitt, (1981, 1988); Gutierrez, Carroll, Wang, Lee, and

Taylor (1995); Lin and Dayton (1997); Lo, Mendell, and Rubin (2001); Masyn (2001);

Masyn and Brown (2001); Mendell, Finch, and Thode (1993); Mendell, Thode, and Finch

(1991); Roeder and Wasserman (1997); Wolfe (1971); and Yang (1999). Most of these

studies concentrated on the distributional properties of the LRT as the criterion for

assessing the number of latent classes (or "components," as it is often referred to in the

literature) in mixture distributions. Studies that are relevant to the current study are

summarized in Table 1.

Wolfe (1971), for example, investigated the utility of the LRT in the latent class

model to reject a mixture of $k_1 = 2$ latent classes from a true $k_0 = 1$ population distribution

using Monte Carlo simulation. Results from his study suggested that under the null

hypothesis of a single population distribution, the LRT, multiplied by a correction factor,

C, would be distributed as χ^2 with degrees of freedom equal to twice the difference in the

number of indicator variables in the model (r), where

$$C = (n - 1 - r - \tfrac{1}{2} k_1) / n,$$

and n indicates the sample size (see McLachlan & Peel, 2001, p. 189). Everitt (1981)

extended this by considering a broader range of sample sizes (Ns = 25, 50, 100, and 200),

number of indicator variables (rs = 5, 10), and replications (200 vs. 500) and concluded

that Wolfe's finding was supported only in cases where sample size was at least 10 times

the number of indicator variables. He further concluded that the power of the LRT was

low; for example, requiring a Mahalanobis distance of 2.5 to 3.0 between latent classes

(i.e., with $N = 200$ and mixing proportions πs = .5) for approximately 80% power. In 1988, Everitt continued his Monte Carlo investigation of the LRT in the latent class model by comparing empirically derived significance levels to 5% and 1% nominal values of the chi-square distribution under the null hypothesis. Results reaffirmed the inadequacy of the LRT under this distribution to reject two-class mixtures from a single population.

Mendell et al. (1991) continued this line of investigation by examining a broader range of experimental conditions consisting of six different mixing proportions (πs = .5 through .95), four sets of standardized differences between means (effects sizes = .5 through 5.0), and three sample sizes (Ns = 25, 50, 100). Data generated for this study, however, were based on a mixture of a two-class distribution, as opposed to the one-class psuedo-populations used by Wolfe (1971) and Everitt (1981, 1988). Critical values for the chi-square distribution under the null hypothesis were based on data from an earlier study (Thode, Finch, & Mendell, 1988). Results supported Everitt's earlier findings that the statistical power of the LRT was (a) strongly related to standardized differences in means between classes, (b) not strongly related to the various mixing proportions, and (c) relatively weak overall (e.g., requiring an effect size of 3.0 and a sample size of 50 for 59% power). Additional analyses indicated that a noncentral chi-square distribution (with 2 degrees of freedom) might be an attractive alternative to the central chi-square distribution for identifying the correct number of latent classes. In turn, Mendell et al. (1993) compared several different model selection criteria (see Table 1) across similar experimental conditions and concluded that, although overall power was weak, the LRT

warranted use in practical applications (outperforming other criteria for mixing proportions between .65 and .80).

As part of their investigation of tomato root initiation, Gutierrez, Carroll, Wang, Lee, and Taylor (1995) conducted a Monte Carlo simulation to examine the ability of the LRT to reject mixtures in a highly skewed one-class population distribution. Results of their study indicated that under the true null hypothesis, the LRT might be distributed as chi-square with between four and six degrees of freedom. Additional results supported earlier findings that the LRT exhibited weak power. For example, the authors showed that a sample size of 350 cases was required for 50% power (at $\pi = .90$, number of replications = 1,000, and a standardized difference between means of approximately 3.0).

More recently, Lo et al. (2001) derived estimates of statistical power for the LRT to reject (a) a false one-class distribution from a mixture of a true two-class pseudo-population and (b) a false two-class distribution from a mixture of a true three-class pseudo-population. Experimental conditions for the one- versus two-class comparison consisted of three mixing proportions (πs = .5, .7, and .9), three sample sizes (Ns = 50, 100, and 200), and three standardized differences between means (ds = 1.0, 2.0, and 3.0). Experimental conditions for the two- versus three-class comparison consisted of two sets of mixing proportions ($\pi_1 = .3$, $\pi_2 = .4$, and $\pi_3 = .3$; and $\pi_1 = .5$, $\pi_2 = .4$, and $\pi_3 = .1$), three sample sizes (Ns = 50, 100, and 200), and four standardized differences between means (ds = 1.0, 2.0, 3.0, and 4.0). One thousand replications were used for both comparisons. The authors examined the power of the LRT distributed asymptotically as the weighted sum of $p + q$ independent chi-square distributions, where $p = 3k_1 - 1$, $q = 3k_0 - 1$, $k_0 = $ the number of classes under the null hypothesis, and $k_1 = $ the number of classes

under the alternative hypothesis (with $k_1 > k_0$). In order to minimize the discrepancy between empirically observed and nominal significance levels, the authors also examined an adjusted LRT:

$$LRT^* = LRT / [1 + \{(p - q) \, ln \, n\}^{-1}].$$

Results supported the earlier work done by Mendell et al. (1991) in that power was primarily influenced by the distance between latent class means and was not strongly associated with different mixing proportions. For example, altering π from .5 to .9 in the two-class model with $N = 200$, $d = 2.0$, and $\alpha = .05$, only increased power by 15% (from .379 to .435). Adequate statistical power for both LRT and LRT* was obtained only when the differences between latent classes (d) was very large. For example, to correctly reject a one-class distribution from a two-class population, $d = 3.0$ was needed for 91.0% and 88.4% power for the LRT and LRT*, respectively (i.e., at $\pi = .5$, $N = 100$, and $\alpha = .05$). To correctly reject a two-class distribution from a three-class population, a difference of three standard deviation units between classes was necessary to yield only 65.3% and 59.8% power, respectively (i.e., at $\pi_1 = .3$, $\pi_2 = .4$, and $\pi_3 = .3$; $N = 100$, and $\alpha = .05$).

Additional studies using Monte Carlo simulation have included penalized information criteria in their investigations. For example, Lin and Dayton (1997) examined the performance of various model selection criteria (AIC, BIC, and Bozdogan's 1987 Consistent AIC) in the latent class model (LCM). Specifically, the authors used latent class models consisting of four dependent indicator variables and equal mixing proportions among latent classes. Experimental conditions included sample size (Ns = 240, 480, 960, and 1,920 cases), model type (i.e., independence, Proctor

uniform, intrusion-omission error, Goodman, extended Proctor models, and extended

intrusion-omission error models), conditional response probabilities, and error rates.

Monte Carlo simulation was conducted using 500 replications assuming (except for the

independence model) a true alternative hypothesis of a two-class mixture distribution in

the pseudo-population. Accuracy was defined as "the percentage of correct selections of

the true model within a set of replications" (Lin & Dayton, 1997, p. 255). Results of the

study indicated that the accuracy of the model selection criteria was related to model type

and sample size with more complex models and smaller sample sizes favoring use of the

AIC over the BIC and CAIC. In general, the AIC indicated a preference to over-

parameterize models, whereas the BIC and CAIC favored models with smaller numbers

of latent classes. Although no single criterion indicated overall superiority, in terms of

applicability, the authors suggested that sample size requirements for the practical use of

the BIC and CAIC may be prohibitive, thus favoring the general use of the AIC.

Yang (1999) also examined the performance of various model selection criteria to

correctly identify mixtures in the LCM; however, he further generalized Lin and

Dayton's (1997) study by examining the ability of several additional criteria, including

the LRT (see Table 1). Yang also considered multiple hypotheses of a true one-class

model tested against false two- through four-class models, a true two-class model tested

against false one-, three-, and four-class models, and so on. Again, Monte Carlo

simulation utilizing 500 replications was used, with equal mixing proportions among all

examined models. Experimental conditions in this study consisted of sample size (Ns =

400 and 800), number of indicator variables (7 vs. 14), and conditional probability

structure (i.e., how distinct the latent classes were from each other, classified as "easy,"

"moderate," "difficult," and "ordered"). Accuracy rates were defined as the percentage of replications that identified the correct number of latent classes across all latent class model comparisons. Results by experimental condition indicated that accuracy was related to the structure of the conditional probabilities and the number of indicator variables within a model. More complex models and models with fewer indicator variables demonstrated lower accuracy rates, except for the AIC, which did not experience a change in accuracy between models with 7 or 14 indicator variables. Larger sample sizes had the effect of improving accuracy for most of the examined model selection criteria (especially for the BIC and ABIC); however, for the AIC and LRT, larger sample sizes were found to produce lower accuracy rates. Results averaged across all experimental conditions indicated that the DBIC and the ABIC exhibited the best accuracy, with the ABIC selecting the greatest overall percentage of true latent classes (approximately 96%). The AIC also performed relatively well, selecting the correct number of latent classes approximately 86% of the time; however this measure again showed a predisposition to favor models with larger numbers of latent classes. On the other hand, the BIC and LRT performed "relatively poorly" with overall accuracy rates of 75% and 73%, respectively, and were not recommended for use in practical research applications.

In tandem Monte Carlo simulation studies, Masyn and Brown (2001) extended the literature from latent class enumeration in the LCM to latent *trajectory* class enumeration in the growth mixture model (GMM). The first focus of this joint study was to analyze the performance of penalized information criteria (i.e., the AIC, BIC, ABIC); classification-based information criteria including the Classification Likelihood Criterion

(CLC; Biernacki & Govaert, 1997), the Normalized Entropy Criterion (NEC; Celeux &

Soromenho, 1996), and the Integrated Classification Likelihood (ICL-BIC; Biernacki,

Celeux, & Govaert, 1998); and the LRT (based on both central and noncentral chi-square

distributions) as they related to the ability to reject over-parameterized models (i.e.,

erroneous multiple latent trajectory classes). Given this focus, this part of the study was

based on a true null hypothesis of a single latent trajectory class for the pseudo-

population versus a false alternative hypothesis of two latent trajectory classes.

Experimental conditions were specified as various sample sizes (Ns = 100, 200), number

of repeated measures (i.e., 3 and 5 time-points), proportion of explained variance (i.e.,

.10, .40, and .80), intercept variances (σs = .25, 1.0, and 2.0), slope means (μs = 0.1, 0.5,

and 1.0), and mixing proportions (i.e., in the alternative two-class models; πs = .10, .20,

and .50). Performance was defined as the percentage of 200 replications that correctly

identified the one latent trajectory class model as having the optimal fit. Results indicated

that the LRT based on an empirically based noncentral chi-square distribution

consistently demonstrated superior ability to identify the true one-class model over the

false two-class model. Moreover, the ICL-BIC and BIC demonstrated "moderate

performance" (in the range of 40% to 60%) across all experimental conditions. The AIC,

BIC, ABIC, and LRT based on a central chi-square distribution, however, showed poor

performance in guarding against over-parameterized models.

The second part of the joint study focused exclusively on the ability of the BIC,

AIC, ABIC, and LRT to correctly determine the number of latent trajectory classes in a

pseudo-population derived from a mixture of two distinct distributions. The comparison

here was between a true alternative hypothesis of two latent trajectory classes (with equal

27

mixing proportions) and false null hypotheses of one- or three-latent trajectory classes (i.e., comparing both under- and over-parameterized models). Experimental conditions included various latent class trajectories expressed as standardized differences in intercept and slope parameters between classes (ds = 0.5, 0.8, and 1.0), proportion of variance explained by level-2 growth parameters (.40, .60 and .80), sample size (Ns = 400, 600, and 800), and number of repeated measures (i.e., 3, 6, and 9 time-points). Here, performance was defined in terms of statistical power, which was operationalized as the proportion of 500 replications that correctly rejected the false null hypotheses. Results contrasted with those found in the first part of the study, with the ABIC and LRT showing superior statistical power and the BIC demonstrating significantly less power. Comparison of results from both parts of the investigation suggested that the utility of model selection criteria was strongly influenced by model over- or under-parameterization. However, the LRT, showed good accuracy overall, and therefore, the authors concluded that the distributional properties of the LRT as a model selection criterion warranted further investigation.

Masyn (2001) expanded the previous Monte Carlo simulation study by considering multiple hypotheses of a (a) true one-class model tested against false two- and three-class models, (b) true two-class model tested against false one- and three-class models, and (c) true three-class model tested against false one- and two-class models (similar to Yang, 1999), as well as including a broader range of latent class trajectories. Results from this study supported the conclusions of the earlier Masyn and Brown (2001) study in that performance of model selection criteria depended on whether the true model was compared to an over-parameterized or under-parameterized model. Specifically, the

noncentral LRT and BIC were superior when over-parameterization was of principal

concern (with average accuracy rates of 99% and 66%, respectively); the LRT, based on

either central or noncentral chi-square distributions, favored rejection of under-

parameterized models (with average accuracy rates of 64% and 65%, respectively); and

all examined model selection criteria performed comparably well when the true two-class

model was compared to false one- and three-class models (with average accuracy rates

ranging from 79% to 82%). Given the superior behavior of the LRT under various

parameterizations, Masyn concluded that its use as an inferential measure of goodness of

fit warranted further research.

The above Monte Carlo studies of latent class enumeration are summarized in

Table 1 (presented in chronological order). Despite somewhat inconsistent methodologies

and ambiguous findings, the Monte Carlo studies described above suggest some general

patterns. First, findings by Dumenci and Windle (2001), Lo et al. (2001), and Mendell et

al. (1991) suggest that distance between latent classes is a primary factor influencing the

ability to correctly identify unobserved heterogeneity; however, various proportions of

cases in mixture distributions are not strongly related to latent class identification.

With regard to examined model selection criteria, Lin and Dayton (1997), Masyn

and Brown (2001), and Yang (1999) found the performance of the BIC to be best when

guarding against over-parameterized or more complex models. That is, the penalty (*r ln

n*) that the BIC imposes upon the estimated log likelihood tends to be conservative in

discouraging latent classes that are not truly present in the data. Conversely, findings

indicate that the AIC typically performs best in models that contain multiple latent

classes and in models with relatively smaller sample sizes. Additionally, the sample size

adjustment to the BIC appears to be an improvement, showing superior performance over the BIC in Masyn (2001), Masyn and Brown (2001), and Yang (1999). Perhaps the most important conclusion gleaned from these studies relates to the utility of the LRT as an inferential method to determine the number of latent classes in mixture distributions. When central chi-square distributions are considered, model selection favors rejection of over-parameterized models. However, utilization of a noncentral chi-square distribution improves the ability of the LRT to reject both false over- and under-parameterized models. Nonetheless, results from all studies point to the fact that accurate detection of latent classes in mixture distributions is not easily obtained and requires a high degree of information (e.g., large number of observations, large separation between classes).

Applications of Growth Mixture Modeling

A review of several applications of growth mixture modeling illustrates the potential of the method as well as inconsistencies in determining the optimal number of latent trajectory classes based on model selection criteria. However, given the recent introduction of the growth mixture model (GMM), the number of studies that have utilized the technique is relatively small (e.g., Colder et al., 2001; Colder, Campbell, Ruel, Richardson, & Flay, 2002; Dedrick, Greenbaum, & Vaughn, 2001; Greenbaum, Dedrick, & Brown, 2000a, 2000b; Greenbaum, Del Boca, Darkes, & Goldman, 2001; Li, Duncan, & Hops, 2001; Muthén, Khoo, Francis, & Boscardin, 2000; Muthén et al., 2002; Oxford et al., 2003; Petras et al., 2001; Tucker, Orlando, & Ellickson, 2003; a summary of these studies is presented in Table 2). The few studies that do exist consistently have relied on the BIC, AIC, ABIC, and/or LRT (along with consideration of substantive theory and model parsimony) to help determine the optimal number of latent trajectory

classes in their sample data. Results from several of these applied studies have shown that enumeration of latent trajectory classes often depends upon which model selection criterion is used.

For example, Petras et al. (2001) used GMM to examine the relationship between school removal (i.e., suspension or expulsion) and developmental trajectories of aggressive behavior in a sample of 200 boys comprising the control condition of a large-scale school-based preventive intervention (Kellam & Reebok, 1992). Results from the study indicated that the optimal number of latent trajectory classes (i.e., best fitting model) for classroom aggression depended on the criterion used—with the BIC indicating a three-class solution and the ABIC, AIC, and LRT indicating four latent trajectory classes in the data. Additionally, Greenbaum, Del Boca, Darkes, and Goldman (2001) examined unobserved heterogeneity in longitudinal patterns of alcohol use in 229 college freshmen. Using the GGMM framework, the authors modeled alcohol consumption as a linear function of time (from start of school) with an additional factor accounting for peak consumption occurring during holiday periods. Their study also showed conflicting numbers of latent trajectory classes, with the BIC indicating a four-class model and the AIC and ABIC indicating a five-class model. Li, Duncan, and Hops (2001) also used GGMM to model longitudinal patterns of alcohol use. Based on a sample of 179 middle and high school adolescents, their study employed a piecewise modeling approach comparing trajectories of alcohol use during middle school and high school years. Enumeration of latent trajectory classes relied on the BIC, AIC, and ABIC as the empirical criteria for model selection. Latent trajectory class enumeration based on BIC, AIC, and ABIC followed a pattern similar to the two aforementioned studies with

the BIC indicating an optimal fit for the one-class model and the AIC and ABIC supporting a two-class model (results for three-class model were found to be unstable). Although the authors did not use the LRT in their determination of latent trajectory classes, examination of log likelihoods for the one-class and two-class models provided in the study indicated that the LRT also would have led to the selection of the two-class model as having a significantly ($p < .05$) better fit than the one-class model.

These studies suggest that the different criteria for latent trajectory class enumeration can produce different results. In fact, half of the eight studies reviewed in Table 2 have model selection criteria producing discrepant numbers of classes, with the BIC consistently identifying one less class than the AIC and ABIC. It is also noted that despite the modest sample sizes used by the studies (ranging from 100 to 442 participants), multiple latent trajectory classes are, nevertheless, present in the data. Although these studies did not present enough information to assess their respective power and accuracy to enumeration latent trajectory classes, they do illustrate that the conditions necessary to identify mixture distributions in longitudinal data are practicable.

Chapter Three

Method

As described in Chapter 1, this study used Monte Carlo simulation as the method to derive estimates of statistical power and accuracy for enumeration of latent trajectory classes in the growth mixture model (GMM). Recently, Monte Carlo simulation has been advanced as a technique to evaluate various aspects of structural equation models. For example, Paxton, Curran, Bollen, Kirby, and Chen (2001) provided detailed steps to help guide researchers using Monte Carlo simulation in their analysis of such models. They suggested that researchers consider the following steps:

1. Develop research questions based on theoretical considerations.

2. Correctly specify the hypothesized model.

3. Choose population parameters for the simulated data.

4. Select the conditions of the experiment.

5. Choose software for data simulation and analysis.

6. Execute the simulation.

7. Troubleshoot problems and verify results.

8. Summarize and present the results.

These steps provide a useful guide for the analysis of power and accuracy in latent trajectory class enumeration, as well. With the exception of Step 8 (which is discussed in the next chapter), these steps also provide the organization for this chapter.

Development of Research Questions

As the first step in the development of a valid and useful Monte Carlo simulation study, the depth and breadth of research questions are of fundamental importance. As Paxton et al. (2001) point out, "…the goal is to construct an optimal match between the research question and the experimental design" (p. 291). The research questions proposed in this study (listed in Chapter One) were derived from a practical consideration of researchers' need to know how well the GMM can correctly identify latent trajectory classes. Specifically, this study was motivated by the paucity of knowledge of the study design and population characteristics that are necessary to identify existing mixture distributions in longitudinal data. Although a variety of experimental conditions can be considered to impact detection of latent trajectory classes, this study examined six specific conditions: three design conditions that can be manipulated by the researcher (i.e., sample size, number of repeated measures, and planned missingness) and three population conditions that can be estimated by the researcher (i.e., between-class differences in growth parameters, percentage of variance explained by intercept and slope parameters, and within-class correlation between intercept and slope parameters). Research questions developed in this study, therefore, were designed to address the "devils" that reside in the details of study design and implementation.

However, research questions in this study also were designed to address the "big picture" as well. In his discussion of Monte Carlo simulation design and analysis, Skrondal (2000) pointed out the need to explicate the relationship between experimental conditions and performance criteria in terms of a *meta-model*. In this study, statistical

power and accuracy were theorized to be positively related to conditions that increased the GMM's ability to discern the true number of latent trajectory classes in longitudinal data (e.g., larger sample size, greater number of repeated measures, more explained variance, etc.). The meta-model for experimental conditions used here related an overall measure of separation or distinction between competing models to empirically derived estimates of power and accuracy. This measure was operationalized as a noncentrality parameter (λ^2) based on the distributions of differences in log likelihoods between "false" one-class and "true" two-class models, and was examined in order to provide a measure of various combinations of experimental conditions not easily displayed in two-dimensional representations of statistical power or accuracy. Furthermore, this measure was incorporated to provide a means to estimate power and accuracy for experimental conditions not examined in this study.

Theoretical consideration also was given to the nature of the hypothesis in question. Following the example provided by McLachlan and Peel (2001, p. 185), the hypothesis for the examination of statistical power was:

H_0: $k = k_0$ (where $k_0 = 1$ latent trajectory class)

versus a true alternative hypothesis,

H_A: $k = k_1$ (where $k_1 = 2$ latent trajectory classes).

Additionally, a second hypothesis was specified to address model accuracy:

H_0: $k = k_0$ (where $k_0 = 1$ or 3 latent trajectory classes)

versus a true alternative hypothesis,

H_A: $k = k_1$ (where $k_1 = 2$ latent trajectory classes).

These hypotheses addressed both the statistical power of the GMM where the researcher wishes to correctly reject models that do not identify multiple latent trajectory classes as well as potential overparameterization of the GMM where the researcher wishes to avoid selecting a model with a greater number of latent trajectory classes than truly exist in the data.

Model Specification

As described in Chapter One, the GMM combines elements of the latent growth curve model and the latent variable mixture model to describe growth curve trajectories for unobserved classes of individuals. Essentially, the model consists of latent growth factors (e.g., intercept [η_0], linear [η_1], or quadratic [η_2] terms) derived from a regression of a repeatedly measured indicator variable (y_{it}). Unobserved heterogeneity in individual growth trajectories is modeled by the addition of mixtures into the model corresponding to *K* latent classes.

Specification of the GMM examined in this study was guided by the desire for results to be generalizable to applied research studies. Although the external validity of a Monte Carlo simulation study is of principal concern (Skrondal, 2000), some simplifying assumptions needed to be made in order for the current study to be practicable. For example, the GMM specified in this study modeled change in normally distributed indicator variables and assumed that individual trajectories were linear (i.e., did not include quadratic or higher order growth factors). Additionally, the model assumed that residual variances among indicator variables were homoscedastic and uncorrelated, and that residual variances and covariances of both indicator variables and growth factors were invariant across latent trajectory classes. Moreover, the GMM examined in this

study did not include covariates or outcome variables (and consequently was not a *general* growth mixture model) nor did it incorporate training data (i.e., cases with known class membership).

Estimation of statistical power and accuracy required the specification of five different types of models. These models are itemized as follows:

1. A one-class GMM estimated from a one-class pseudo-population ($K1_T$).

2. A two-class GMM estimated from a one-class pseudo-population ($K2_F$).

3. A one-class GMM estimated from a two-class pseudo-population ($K1_F$).

4. A two-class GMM estimated from a two-class pseudo-population ($K2_T$).

5. A three-class GMM estimated from a two-class pseudo-population ($K3_F$).

For example, in order to estimate power and accuracy at a nominal $\alpha = .05$, determination of effective Type I error rates for each experimental condition required comparing model selection criteria between $K1_T$ versus $K2_F$ models (for power) and between $K2_T$ versus $K3_F$ models (for accuracy). Similarly, estimation of statistical power to detect a two-class model over a one-class model required the comparison of $K1_F$ versus $K2_T$ models; estimation of accuracy required the comparison of $K2_T$ with both $K1_F$ and $K3_F$ models. See Appendix A for Mplus input syntax used to estimate these models.

Population Parameters

Simulated data were constructed to represent samples drawn from 13 distinct pseudo-populations consisting of a mixture of two latent trajectory classes (i.e., with significantly different between-class intercept and linear slope parameters). Specific values for growth parameters, their variances, and covariances used in creating the 13 pseudo-populations are presented in Table 3. The procedure for generating the simulated

data was as follows: First, a matrix of data elements was created representing individual cases by repeated measurement occasions for each of the two specified latent trajectory classes. Values for the dependent indicator variables were constructed as linear functions of a Level-1 intercept, linear slope, and random error ε_i, $\sim N(0, \sigma^2)$, where σ^2 represents the Level-1 residual variance for each unique population distribution. Specific values for the Level-1 intercept and linear slope parameters were derived from the class-specific Level-2 intercept and linear growth parameters, plus their respective error terms, $\zeta_{0i} \sim N(0, \sigma_0^2)$ and $\zeta_{1i} \sim N(0, \sigma_1^2)$. Specific values for the Level-1 and Level-2 residual variances used to create the simulated data were based on a review of several longitudinal studies that provided information about their variance components (Greenbaum et al., 2000a, 2000b; Greenbaum et al., 2001; Greenbaum & Brown, 2002). Averaging results from these studies, approximately 5% of Level-2 variance was allocated to the linear slope parameter (i.e., $\sigma_1^2 = 3.0$) with the remaining Level-2 variance allocated to variability in intercepts (i.e., $\sigma_0^2 = 57.0$).

Values for Level-2 intercept and linear growth parameters were designed to be different for each of the two latent trajectory classes with values for Class 1 chosen to represent a zero growth class of individuals (i.e., $\eta_{1k1} = 0$) and values for Class 2 chosen to represent a class of individuals with decreasing scores over time (i.e., $\eta_{1k2} < 0$). To avoid negative scores for individuals from Class 2, intercept values were set to 100 for Class 1 (i.e., $\eta_{0k1} = 100$). A graphical example of simulated data generated from population parameters for Population Y is presented as spaghetti plots in Figure 5.

38

Selection of Experimental Conditions

Six specific experimental conditions were tested in this study: (a) standardized between-class differences in intercept and slope parameters, (b) percentage of variance explained by intercept and slope parameters, (c) correlation between intercept and slope parameters, (d) sample size, (e) number of repeated measures, and (f) planned missingness. Overall, these represented 36 unique experimental conditions (i.e., not all conditions were cross-classified; see Table 4 for a complete list of all experimental conditions and corresponding population/design codes). The first three conditions represented characteristics of population distributions among latent trajectory classes; whereas the fourth, fifth, and sixth conditions related to characteristics of research design.

First, between-class differences in intercept and slope parameters were operationalized as the difference in both mean intercept and slope values between latent trajectory classes divided by each parameter's respective (class invariant) standard deviation. These values were expressed as standardized between-class differences (*d*) of 0.30, 0.50, 0.80, 1.00, and 1.20 (for Populations A through E, respectively). Second, the percentage of explained variance was operationalized as the Level-2 variance (sum of the intercept and slope variances, when the two growth parameters were orthogonal) divided by the total variance (i.e., sum of Level-1 and Level-2 variances). Five distinct percentages (i.e., 40%, 50%, 60%, 70%, and 80%) were created by altering the amount of Level-1 variability from a maximum ($\sigma^2 = 90$) in Population F to a minimum ($\sigma^2 = 15$) in Population I (see Table 4). Third, the covariance between Level-2 intercept and slope parameters for each population distribution was varied to represent both positive (*r*s = .30 and .60, see Populations AG and AH, respectively) and negative (*r*s = -.30 and -.60, see

39

Populations AI and AJ, respectively) correlations, as well as zero correlations between

the parameters (i.e., $r = .00$, for all other populations).

Fourth, sample sizes consisted of Ns = 200, 400, 600, 800, and 1,000 participants

evenly divided between the two classes. Fifth, the number of repeated measures consisted

of 3, 4, 5, 6, 9, and 12 evenly spaced time points. And finally, the role of planned

missingness in longitudinal designs was examined. In each of two different designs

(shown in Table 5) represented by planned missingness at (a) every other time point and

(b) every 2nd and 3rd time point, the total amount of missing data was varied in increments

of 25% from 0% (i.e., no missing data) to 100% (i.e., complete missing data for that time

point). Examination of this final condition was included to provide information on the

efficiency of incorporating planned missingness in longitudinal designs. Values for

sample sizes, repeated measures, and planned missingness were selected to represent

reasonable ranges from which longitudinal studies could be designed. Because of the

importance of sample size consideration on the development of research designs, this

experimental condition was examined for three different pseudo-populations consisting

of between-class differences in intercept and slope parameters of ds = .50, .80, and 1.00.

Different values for numbers of repeated measures and planned missingness, however,

were examined for one particular population distribution consisting of a between-class

difference in growth parameters of 1.0 standard deviation units, a percentage of variance

explained by growth parameters of 60%, and no correlation between intercept and slope

growth parameters ($r = .00$). These values were selected to represent a typical distribution

from which estimates of statistical power and accuracy for the various design

characteristics could be readily generalized.

Although not a specific experimental condition, this study also examined statistical power and accuracy as they related to an omnibus measure of the discrepancy between latent trajectory classes. This measure was operationalized as a noncentrality parameter (λ^2) based on a noncentral chi-square distribution and was approximated as

$$\lambda 2 = \frac{\sum\limits_{n=i}^{n} -2\,(logL_{K1_F} - logL_{K2_T})}{n} - df$$

where λ^2 was defined as the average of –2 times the difference in log likelihoods between one-and two-class models across n replications, minus degrees of freedom equal to the difference in the number of freely estimated parameters between the models (i.e., degrees of freedom = 3 for all experimental conditions examined in the study). Larger values for λ^2 correspond to greater separation between latent trajectory classes and were theorized to be associated with greater power and accuracy.

Software

Simulated data for the study were generated using both SAS Interactive Matrix Language (IML; SAS Institute Inc., 1988) and Mplus v.2.1 (Muthén & Muthén, 2001). Monte Carlo simulation in Mplus allows for the flexibility to define pseudo-populations and estimate corresponding models from both (a) data generated from other software programs and (b) data generated from within Mplus, itself (see Masyn & Muthén, 2000; Muthén, 2002; Muthén & Muthén, 2002; and Muthén & Asparouhov, 2002 for detailed procedures regarding Monte Carlo simulation using Mplus). Both procedures were used for this study. For example, the detailed specification of missing data patterns used in examining planned missingness designs was performed using SAS IML, with the

externally generated data read into Mplus for model estimation (see Appendix B for an

example of SAS IML syntax used to generate sample data for Population Y). The

procedure for executing the GMMs and concatenating the resulting output files was

facilitated by the use of several DOS batch files made available through the Mplus

website (www.statmodel.com). See Appendix C for examples of DOS batch files used in

the analysis of simulated data from Population Z. Note that these files were modified,

somewhat, from their original structure for sequential analysis of the one-, two-, and

three-class models tested in this study.

Data generation and model estimation also were conducted completely within

Mplus. Here, attributes for the pseudo-populations (e.g., number of mixture distributions;

mixing proportions; and growth parameter means, variances, and covariances), design

characteristics (e.g., number of repeated measures, sample size, missing data patterns),

and number of replications all can be specified in a single input program (see Appendix

A for examples of Mplus input syntax for analysis of models based on Population Y).

Results from this procedure were automatically concatenated in output files containing

log likelihoods, parameter estimates, variances, covariances, and convergence

information. Program syntax to read-in Mplus output files, construct model selection

criteria, and determine adjusted power and accuracy were written and analyzed in SPSS

for Windows v.10.0 (SPSS Inc., 2000). (See Appendix D for an example of SPSS syntax

used for analysis of Population Y).

Execution of Simulation Runs

As Paxton et al. (2001) has suggested, a successful Monte Carlo simulation study

requires that several decisions be made prior to the execution of the runs. First among

these is determination of the number of replications. In this study, the number of replications for each experimental condition was determined by the desired confidence interval for estimates of statistical power and accuracy. Based on the binomial distribution where $p*$ indicates the proportion of replications that identifies the correct two-class model and $q*$ indicates the proportion of replications that identifies the erroneous one- or three-class models, the variance of $p*$ is given by

$$Var(p*) = p* \, q* \, / \, n.$$

Alternatively, the number of replications n is given by

$$n = p* \, q* \, / \, Var(p*).$$

Table 6 shows the number of replications required for a 95% confidence interval at $p* = .90$ and $p* = .50$. As shown in the table, 400 replications were required for 95% confidence level of $\pm .05$ at $p* = .50$ and $\pm .03$ at $p* = .90$. For this study, 400 replications would yield an acceptable level of precision for estimates of statistical power. However, to account for models that failed to reach convergence, a total of 500 replications were conducted for each experimental condition.

Other important issues regarding the execution of simulation runs in this study relate to the identification of starting values and model convergence. All replications conducted in this study relied upon a single set of starting values consisting of (a) sample means for intercept and slope parameters in one-class models, (b) ± 0.5 standard deviation units from sample means in two-class models, and both (a) and (b) for three-class models. Additionally, an arbitrary maximum of 5,000 iterations was allowed for convergence of the GMMs. Models that did not converge within this limit were considered failures and were not eligible for selecting the best fitting model (although

43

they were retained in the denominator for calculation of power and accuracy rates). Based on this criterion, a small percentage (0.7%) of replications did not converge (convergence rates for each of the five types of models are shown in Table 7 by population/design).

As mixture models have been "known to sometimes generate a likelihood function with several local maxima" (Muthén & Muthén, 2001, p. 292), it would be preferable to reanalyze all nonconvergent models with alternate starting values. However, given that this study required 90,000 separate analyses (i.e., 36 experimental conditions X 5 models X 500 replications), this approach was impracticable. Alternatively, a random sample of 5% of nonconvergent replications for the $K3_F$ model were inspected to determine if using different (i.e., more disparate) starting values would result in different convergence rates. Altering starting values to ± 1.0 standard deviation resulted in almost identical rates of convergence. Moreover, *t* tests of log likelihoods values revealed no significant difference (all *p*s > .05) between mean values for replications that converged versus those that did not converge (log likelihood values were available for both conditions). Based on this, the original decision rule regarding nonconvergent replications as "failures" was retained.

Verification

To insure the validity of study results, frequency distributions and box plots of estimated parameters were visually inspected for questionable values (e.g., negative variances, extreme values for log likelihoods). Additionally, results from a random selection of 2% of replications within each experimental condition were visually double-checked to insure that model selection criteria were correctly matched across the five models.

44

Chapter Four

Results

Results of the Monte Carlo simulations begin with a description of the derivation

of effective Type I error rates and their corresponding critical values used in establishing

a common $\alpha = .05$ across all examined model selection criteria. In turn, results of

simulations for each of the six experimental conditions are reported individually. This is

followed by a summary of estimates of statistical power and accuracy across all 36

combinations of experimental conditions, as operationalized by an omnibus measure of

the degree of separation between models (λ^2). Finally, the chapter concludes with results

of a motivating example, which field test the derived values of λ^2 generated from the

Monte Carlo simulations to actual data from an evaluation of a longitudinal service

delivery demonstration project.

Effective Type I Error Rates and Critical Values

Effective Type I error rates for estimates of statistical power and accuracy were

derived from empirical distributions of the differences in AIC, BIC, and ABIC values

comparing (a) one-class models estimated from one-class pseudo-populations ($K1_T$)

versus two-class models estimated from one-class pseudo-populations ($K2_F$), and (b)

two-class models estimated from two-class pseudo-populations ($K2_T$) versus three-class

models estimated from two-class pseudo-populations ($K3_F$). These comparisons assumed

the traditional minimal-value rule-of-thumb for penalized model selection criteria to

assess optimal model fit. Specifically, effective Type I error rates for estimates of statistical power in two- and three-class models were operationalized as the proportion of the 500 replications that resulted in a positive value for $K1_T$ - $K2_F$ and $K2_T$ - $K3_F$ model comparisons, respectively. Larger values for these rates indicated a greater probability of falsely rejecting the "true" one- and two-class models, respectively. Using data from Population K as an example, 14.4% of replications resulted in a greater ABIC value for the $K1_T$ model than the comparable $K2_F$ model—a Type I error rate almost three times greater than expected at α = .05. Figure 6 provides a graphical representation of this example. The histogram shown in this figure represents the distribution of $K1_T$ - $K2_F$ values for the ABIC, with positive values (i.e., to the right of zero) indicating better fit for the erroneous $K2_F$ model. In contrast, the effective Type I error rate for the BIC in Population K was very conservative (.002 for comparison of $K1_T$ vs. $K2_F$ models), indicating that the BIC rarely demonstrated better fit for the "false" two-class model.

Effective Type I error rates for the LRT were derived from -2 times the difference in log likelihoods between $K1_T$ versus $K2_F$ models and $K2_T$ versus $K3_F$ models, respectively. These differences were evaluated using a chi-square distribution with degrees of freedom equal to the difference in the number of freely estimated parameters between competing models, which for all examined conditions was χ^2 (3, N = 500) = 7.81. For Population K, 7% of replications resulted in log likelihood values greater than 7.81 for $K1_T$ - $K2_F$ model comparisons, a value only slightly larger than the target Type I error rate of .05.

Effective Type I error rates, averaged within and across experimental condition and quartiles of λ^2, are shown in Table 8 (for $K1_T$ vs. $K2_F$) and Table 9 (for $K2_T$ vs. $K3_F$).

46

Although little variation was found among experimental conditions or quartiles of λ^2,

effective Type I error rates did vary substantially by model selection criterion. Overall,

the ABIC demonstrated effective Type I error rates closest to the desired $\alpha = .05$,

averaging .046 for $K1_T$ versus $K2_F$ and .062 for $K2_T$ versus $K3_F$, across all experimental

conditions. This was followed by the LRT, which demonstrated slightly more liberal

error rates of .082 and .105, respectively. The AIC demonstrated the most liberal error

rates of the four criteria (.167 and .213, respectively), suggesting that use of the minimal-

value rule-of-thumb for this criterion would over-select models with larger numbers of

latent trajectory classes. Alternatively, the BIC demonstrated very conservative error

rates overall (.001 for both model comparisons), suggesting that the BIC would over-

select models with smaller numbers of latent trajectory classes.

In order to fairly compare the statistical power and accuracy of the model

selection criteria at a common Type I error rate, critical values were obtained at the 95th

percentile of the above described distributions, corresponding to a nominal $\alpha = .05$.

Empirically derived critical values for each experimental condition are shown in Table 10

and denote the point in a distribution at which 95% of the differences between competing

models for each criterion is smaller than the identified critical value. These values can be

interpreted as the minimum degree of separation between competing models that would

be necessary to correctly identify the more complex $(k_0 + 1)$ model at an effective Type I

error rate of .05. Again, using Population K as an example, the 95% critical value for the

comparison of ABIC values between $K1_T$ versus $K2_F$ models is 2.32, suggesting that the

ABIC value for a two latent trajectory class model be at least this amount less than the

ABIC value for a one-class model to demonstrate significantly better fit at $\alpha = .05$ (see

47

Figure 6). By comparison, for the same population, the value of the BIC for the two latent

trajectory class model could be up to 7.18 *greater* than the corresponding value for the

one-class model for the two-class model to demonstrate significantly better fit at α = .05.

These critical values were used to derive estimates of statistical power and

accuracy at a common Type I error rate (α = .05). These are referred to as *adjusted* power

and *adjusted* accuracy, respectively. For the ABIC, AIC, and LRT, adjusted estimates of

power and accuracy based on the empirically derived critical values are lower than the

corresponding rates based on the minimal-value rule-of-thumb. For the BIC, however,

adjusted power and accuracy rates are higher than estimates derived by taking the lowest

BIC value across competing models. Not surprisingly, use of the empirically derived

critical values to estimate power and accuracy at a common .05 Type I error rate resulted

in identical power and accuracy estimates (for a particular experimental condition) across

all model selection criteria; that is, the four criteria are essentially equated and, therefore,

only one curve (each) is needed to describe adjusted power and accuracy vis-à-vis the

various experimental conditions.

Standardized Between-Class Differences in Growth Parameters

Estimates of adjusted statistical power to correctly identify the "true" two-class

(K2$_T$) model over the "false" one-class (K1$_F$) model are shown in Figure 7 for five

selected values of standardized between-class differences in intercept and slope

parameters (*d*). Estimates shown in this figure assumed an overall sample size (*N*) of 600,

equally spaced across six time points, no missing data, zero correlation between intercept

and slope parameters (*r* = 0), and 60% of total variance explained by intercept and slope

growth parameters. These conditions represent Populations A through E for *d*s = .30, .50,

.80, 1.00, and 1.20, respectively (see Table 4). As shown in Figure 7, when Class 1 and

Class 2 intercept and slope parameters are .30 standard deviation units apart (d = .30),

adjusted power is very low (.052). However, adjusted power increases rapidly with larger

d values, reaching an estimated .986 for d = 1.20. Estimates of adjusted accuracy

corresponded very closely with adjusted power estimates, ranging from .050 for d = .30

to .922 for d = 1.20. Estimates of adjusted power and accuracy for Populations A through

E are presented in Table 11 (ordered by λ^2).

Percentage of Variance Explained by Growth Parameters

Figure 8 shows estimates of adjusted power by the percentage of variance

explained by intercept and slope parameters. Estimates shown in this figure are based on

an overall sample size of 600, six time points, no missing data, zero correlation between

intercept and slope parameters, and ds = 1.00. These conditions consist of Populations F,

G, D, H, and I for percentages = 40%, 50%, 60%, 70%, and 80%, respectively (see Table

4). Here, power was positively and linearly related to greater percentages of explained

variance, doubling in magnitude from 40% (adjusted power = .476) to 80% (adjusted

power = .950). Again, estimates of adjusted accuracy were close to estimates of adjusted

power and followed a similar linearly increasing pattern. Estimates of adjusted accuracy

increased from .444 for 40% explained variance to .890 for 80% explained variance.

Estimates of adjusted power and accuracy for Populations F, G, D, H, and I are presented

in Table 11.

Correlation Between Intercepts and Slopes

Figure 9 displays the adjusted power curve estimated as a function of the

correlation between intercept and slope growth parameters (r). Estimates shown here are

based on $N = 600$, six time points, no missing data, $d = 1.0$, and 60% variance explained by growth parameters. These conditions represent Populations AF through AJ for rs = -.60, -.30, .00, .30, and .60, respectively (see Table 4). Results indicated that adjusted power and accuracy were inversely related to greater positive correlations between intercepts and slopes, with a disproportionate reduction in power for more negative correlations; for example, from a power estimate of .732 (at $r = -.60$) to .180 (at $r = -.30$). Estimates of adjusted accuracy were approximately 13% smaller than adjusted power estimates. Estimates of adjusted power and accuracy for Populations AF through AJ are presented in Table 11.

Sample Size

Estimates of adjusted power are shown in Figure 10 as three separate curves for ds = 0.50, 0.80, and 1.00, respectively. Each curve shows the increase in adjusted power as sample size increases in multiples of 200 from $N = 200$ to $N = 1000$. Estimates shown here are similarly based on six time points, no missing data, rs = .00, and 60% variance explained by growth parameters. These conditions represent Populations B, C, D, and J through T (see Table 4). The three curves depicted in the figure show that increasing sample size has a greater effect on adjusted power when $d = 1.00$ (from power = .318 at $N = 200$ to .952 at $N = 1000$), as opposed to $d = .80$ (with power increasing from .120 to .624, respectively) and $d = .50$ (from .060 to .112 power, respectively). Ancillary analyses indicated that increasing sample size had no major effect on power when $d = 0.30$. Estimates for adjusted accuracy for the various sample sizes were approximately 6% to 7% less than estimates of adjusted power (see Table 11).

Number of Repeated Measures

Estimates of adjusted power for 3, 4, 5, 6, 9, and 12 time points are shown in Figure 11. These estimates are based on $N = 600$, $r = .00$, $d = 1.00$, 60% of variance explained by growth parameters, and no missing data. These conditions represent (in order) Populations U, V, W, D, X, and Y (see Table 4). A substantial increase in adjusted power occurred when the number of repeated measures was increased from 3 time points to 6 time points (from .084 to .824, respectively). The increase in adjusted power from 6 to 12 time points, however, was disproportionately smaller (from .824 to .954, respectively). Again, estimates of adjusted accuracy corresponded very closely with adjusted power estimates, ranging from .074 at 3 time points to .878 at 12 time points. Estimates of adjusted power and accuracy for the above populations are presented in Table 11.

Planned Missingness

Figure 12 shows the decrement in adjusted power for two different patterns of data missing by design. The first design incorporated planned missingness at every other time point and the second design incorporated planned missingness at every second and third time point (see Table 5). For each respective design, the percentage of planned missingness in selected time points was increased by increments of 25% from no missing data (corresponding to a 12 time point repeated measures design) to 100% missing data (corresponding to six time points for Design 1 and four time points for Design 2). Populations for these experimental conditions were based on $N = 600$, $r = .00$, $d = 1.00$, 60% of variance explained by growth parameters, and 12 time points. These conditions represent Populations Y, Z, AA, AB, and D for Design 1 and Populations Y, AC, AD,

51

AE, and V for Design 2. For Design 1, the drop in adjusted power was only slight,
decreasing from .954 at no missing data to .824 when 100% of the cases were missing
data at every other time point. However, for Design #2, the drop in power was more
substantial, decreasing from .954 at 0% missing data to .334 when 100% of the cases
were missing data at every second and third time point. Estimates of adjusted power and
accuracy for the above populations are presented in Table 11.

Degree of Separation Between $K1_F$ and $K2_T$ Models (λ^2)

To generalize across combinations of experimental conditions not examined in
this study, estimates of adjusted power and accuracy were related to an omnibus measure
of the degree of separation between a "false" one-class ($K1_F$) and "true" two-class ($K2_T$)
models (denoted as λ^2). Again, λ^2 is defined as

$$\lambda^2 = \frac{\sum_{n=i}^{n} -2\,(logL_{K1_F} - logL_{K2_T})}{n} - df$$

that is, as the average of –2 times the difference in log likelihoods between the competing
models across *n* replications, minus the difference in the number of freely estimated
parameters between the models (i.e., df = 3 for all model comparisons). Estimated values
of λ^2 for the 36 experimental conditions examined in this study are presented in Table 11,
where larger values of λ^2 represent greater overall separation between the two models.

Regressing the derived estimates of adjusted power and accuracy on λ^2 provided a
basis by which power and accuracy could be extrapolated for all values of λ^2. Results
from these regressions indicated that adjusted power and accuracy were accurately

52

estimated by regression equations including linear and quadratic terms. Specifically, adjusted power was best modeled as

$$\hat{Y}_{\text{adjusted power}} = -.051 + .098(\lambda^2) - .002(\lambda^2)^2$$

resulting in an $R^2 = .99$ and a model standard error = .042. Similarly, adjusted accuracy was best modeled as

$$\hat{Y}_{\text{adjusted accuracy}} = -.062 + .092(\lambda^2) - .002(\lambda^2)^2$$

with a corresponding $R^2 = .98$ and model standard error = .048. These relationships are shown graphically as adjusted power and accuracy curves by λ^2 (see Figure 13). As shown in the figure, a λ^2 value of 12.35 is needed to achieve adjusted power of .80 (for detection of $K2_T$ relative to $K1_F$). For an adjusted accuracy of .80, $\lambda^2 = 14.54$ is required (for the same model comparison).

Motivating Example

A motivating example of power estimation for latent trajectory class enumeration in the growth mixture model is provided to illustrate the procedure. Data used in this example were provided by the Comprehensive Community Mental Health Services for Children and Their Families Program (ORC Macro International, 1995), a longitudinal *system-of-care* demonstration project that evaluated an integrated systems approach for the provision of children's mental health services. As part of the evaluation, a series of growth mixture models were estimated to test for between-group differences (i.e., intervention versus comparison sites) in trajectories of delinquent behavior. The sample for the evaluation consisted of 442 children measured at intake into service and every six months thereafter for a two-year period. Measures of delinquent behavior consisted of Child Behavior Checklist/4-18 delinquency syndrome T-scores (Achenbach, 1991).

Approximately 10% of the sample was missing data at the second measurement occasion, 11% at Time 3, 19% at Time 4, and 35% at Time 5. As noted earlier, results based on the entire sample indicated that a four-class growth mixture model exhibited optimal fit to the data (Greenbaum & Brown, 2002). However, for simplicity, the motivating example is restricted to an analysis of linear trajectories only, and does not incorporate results with quadratic growth trajectories nor adjustment for propensity scores, both of which were included in the original evaluation.

Model selection information for fitted one-, two-, and three-class linear growth mixture models, based on data from the evaluation, are shown in the top half of Table 12. Although the BIC indicated a preference for the one-class model over the two-class model, values for the AIC, ABIC, and LRT identified the two-class model as having better fit to the data than a one-class model. Furthermore, all model selection criteria identified the three-class model as having best overall fit, compared to both one- and two-class models (a four-class linear model did not converge to a stable solution). Latent trajectory classes for the three-class model consisted of (a) "low-declining," (b) "high-declining," and (c) "high-increasing" groups, comprising approximately 44%, 51%, and 5% of participants, respectively.

Estimates of statistical power to reject the one-class model in favor of a two-class model, and the two-class model in favor of a three-class model, were based on 500 replications of simulated data using parameter estimates, mixing proportions, and a missing data pattern obtained from the analysis of the evaluation data (see Table 13 for parameter estimates for one-, two-, and three-class growth mixture models). Starting values for growth parameters in the Monte Carlo simulation models were based on mean

54

values from the one-class (i.e., latent growth) model, and deviations of ±0.5 standard

deviation units for the two-class model and ±1.0 standard deviation units for the three-

class model. Variances and covariances for estimated parameters were constrained to be

equal across classes for both two- and three-class models. All Monte Carlo simulation

analyses were performed using Mplus Monte Carlo simulation procedures (Masyn &

Muthén, 2000; Muthén, 2002; Muthén & Asparouhov, 2002; Muthén & Muthén, 2002).

See Appendix E for Mplus input programs used for the motivating example.

Values for the AIC, BIC, ABIC, and log likelihoods averaged across 500

replications of simulated "true" one-, two-, and three-class models are shown in the

bottom half of Table 12. Mean values for the penalized information criteria and log

likelihoods derived from Monte Carlo simulation corresponded very closely to values

generated from the analysis of the original evaluation data, which is to be expected given

that the parameters for the pseudo-populations in the Monte Carlo simulations were

derived directly from the analysis of the original evaluation data. Minimal values for all

model selection criteria similarly identified a three-class linear growth mixture model as

having optimal fit.

Statistical power to detect a "true" two-class model ($K2_T$, defined as estimation of

a two-class model drawn from a two-class pseudo-population) was estimated at an

effective Type I error rate of .05. To accomplish this, critical values for each model

selection criterion were derived from empirical distributions of differences between

"false" two-class models ($K2_F$, i.e., a one-class pseudo-population for estimation of a

two-class model) and "true" one-class models ($K1_T$, i.e., a one-class pseudo-population

for estimation of a one-class model). For example, the distribution of the difference in

BIC values between simulated $K2_F$ and $K1_T$ models resulted in a value of -8.76 for the

95th percentile, suggesting that the BIC for a two-class model could be up to 8.76 points

greater than the BIC for the corresponding one-class model and still demonstrate

significantly better fit at $\alpha = .05$. Statistical power for a "true" three-class model ($K3_T$,

i.e., estimation of a three-class model drawn from a three-class pseudo-population) also

was estimated at an effective Type I error rate of .05, with critical values similarly

obtained as the result of comparisons in the empirical distributions of model selection

criteria between $K3_F$ (i.e., two-class pseudo-population for estimation of a three-class

model) and $K2_T$ models.

Statistical power for the "true" two-class model was defined as the proportion of

500 replications that contained a significantly better fit for the $K2_T$ model relative to the

$K1_F$ model (i.e., a two-class pseudo-population for estimation of a one-class model),

relying on the empirically derived critical values to achieve an effective Type I error rate

of .05. Likewise, statistical power for the "true" three-class model was defined as the

proportion of replications that significantly identified the $K3_T$ model as having better fit

to the data than a competing model that estimated two latent trajectory classes from a

three-class pseudo-population (denoted as $K2_{FF}$). (Four replications for the $K3_T$ model

resulted in unstable parameter estimates and consequently were rejected in favor of the

competing $K2_{FF}$ model.) Comparison of model selection criteria for $K1_F$ and $K2_T$ models

indicated that 89.4% of replications identified the $K2_T$ model as having significantly

better fit than the $K1_F$ model. For the comparison of $K2_{FF}$ and $K3_T$ models, 91.4% of

replications identified the $K3_T$ model as having significantly better fit than the $K2_{FF}$

model. Therefore, based on data pattered after the demonstration project evaluation, the

56

Monte Carlo simulations resulted in estimates of .894 power to detect a "true" two-class linear growth mixture model over a one-class model, and .914 power to detect a "true" three class model over a two-class model.

These estimates were related back to the functional relationship between λ^2 and estimates of statistical power derived across the 36 experimental conditions described earlier. Values of λ^2 corresponding to the estimates of power from the simulation of evaluation data were 18.88 and 34.49, respectively. Inserting these values into the equation described earlier resulted in predicted power estimates of .957 (95% confidence interval = .874 to 1.00) and 1.00 (95% confidence interval = .917 to 1.00), respectively.

Chapter Five

Discussion

This study was guided by a desire to provide a basis for applied researchers to make design decisions to achieve a desired level of power and accuracy for determining the optimal number of latent trajectory classes. However, to accomplish this requires several steps, coinciding with the multiple objectives of this study. In this chapter, the discussion of findings from this study begins with a description of the general procedure for conducting the Monte Carlo simulations for latent trajectory class enumeration. This step-by-step process is described in order to facilitate power and accuracy estimation for researchers who would like to produce estimates for an existing or proposed study of their own, or who would like to extend research related to latent trajectory class enumeration into new areas (e.g., other experimental conditions not examined in this study). The discussion then continues with a description of the specific results of the study as they relate to the initial research questions posed in Chapter One. Susbsequently, as examples of how the procedure for power estimation is integrated with the results derived from the Monte Carlo simulations, two hypothetical applications are presented. In turn, this leads to a discussion of findings from the motivating example using real data from the evaluation of a longitudinal demonstration project. This is followed by a description of limitations to the current study, which naturally flows into possible

58

directions for future research. Finally, the chapter concludes with several general conclusions gleaned from the study.

Summary of Monte Carlo Simulation Procedure

The procedure used in this study involved Monte Carlo simulation of data, which are based on a particular set of design and population characteristics, and penalized information criteria (i.e., the AIC, BIC, and ABIC) or log likelihood values (i.e., the LRT) to develop the decision rules for enumerating latent trajectory classes. Although the procedure is relatively straightforward, a number of detailed decisions need to be made. The steps a researcher has to go through to estimate statistical power or accuracy using Monte Carlo simulation are documented as follows.

First, using information from previous studies or, in the absence of such studies, developing values based on theory, the researcher needs to specify the characteristics of the pseudo-population from which simulated sample data will be drawn. These should include but not be limited to: (a) distribution of the dependent indicator variable (y_{it}), (b) the hypothesized number of latent trajectory classes in the population, (c) degree of separation between latent trajectory classes, (d) functional form of trajectory classes (e.g., linear, quadratic, etc.), (e) proportion of cases in each latent trajectory class (i.e., mixing proportion), (f) covariation among growth parameters, and (g) variability within and between latent trajectory classes (e.g., decomposition of variance components among growth parameters, invariance or noninvariance among classes). Second, the researcher should similarly identify the characteristics of the particular longitudinal design of interest, including, for example, (a) the relevant sample size, (b) number and spacing of repeated measures, and (c) degree and pattern of missing data. Values for these

population and design characteristics are specified in computer input programs, as shown in Appendix A. Third, consideration should be given to the mechanics of conducting the Monte Carlo simulation runs. For example, the researcher needs to decide on the acceptable degree of precision of resultant estimates of statistical power and accuracy. This translates into determining the number of replications for simulated data, and the maximum allowable number of iterations in the estimation process. In this study, 500 replications were determined to be sufficient for 95% confidence intervals of no greater than $\pm .045$ around estimates of statistical power and accuracy. Attention also needs to be given in specifying the correct random number seeds in generating data for comparisons of different models within a common data set.

Fourth, the research hypothesis or hypotheses must be made explicit and also should be accurately represented in the corresponding computer input programs. For example, if the goal of the Monte Carlo simulation is to ascertain estimates of power to detect a mixture of two latent trajectory classes, as opposed to a single underlying distribution in the population, the null hypothesis represents the $k_0 = 1$ situation and the alternative hypothesis represents the $k_1 = 2$ situation. This translates into a comparison of competing hypotheses: the first (null) hypothesis consists of estimating of a "false" one-class model from a "true" two-class pseudo-population; the second (alternative) hypothesis consists of estimating a "true" two-class model from the "true" two-class pseudo-population (see Appendix A for an example of Mplus input syntax for this comparison). If the researcher additionally wishes to rule out the possibility of a three-class solution (i.e., to determine the "accuracy" of the two-class model), the null hypothesis becomes $k_0 = 1$ or 3, and the alternative hypothesis is modified to include a

"false" three-class model estimated from the "true" two-class population (again, see Appendix A for an example).

Fifth, the researcher needs to decide on the criteria for selecting the optimal model. Implicit in the decision rules the researcher chooses in model selection is determination of the Type I error rate. If the traditional minimal-value-across-competing-models rule-of-thumb is used with the AIC, BIC, or ABIC, then the researcher needs to be cognizant of effective Type I error rates associated with each criterion. Results of this study support the use of the ABIC, which consistently demonstrated Type I error rates closest to the conventional $\alpha = .05$ (see response to research Question #7 below). Alternatively, as was done in this study, the researcher may choose to use empirically derived critical values for the AIC, BIC, ABIC, or LRT in choosing the optimal model. Here, an additional model comparison is required consisting of a "true" one-class model estimated from a "true" one-class pseudo-population (the null hypothesis) versus a "false" two-class model estimated from a "true" one-class pseudo-population (the alternative hypothesis). The difference in model selection criterion values at the 95th percentile of the distribution identifies the critical value that should be used to compare values of the model selection criterion in estimating power (at a nominal $\alpha = .05$). In this way, differences in Type I error rates are equated by the different critical values and the decision of which model selection criterion to use is not important—all criteria will result in identical estimates of power and accuracy.

Finally, results from the various model comparisons are linked by replication number and the proportion of replications that (correctly) reject the null hypothesis is found using the decision rules just described. An example of computer programming that

accomplishes this can be seen in Appendix D. This procedure can be repeated with

varying values of design or population characteristics to determine the specific set of

characteristics necessary for a particular level of power or accuracy.

Summary of Findings for the Research Questions

The goal of this study was to produce estimates of statistical power and accuracy

for six specific experimental conditions chosen to represent design and population

characteristics commonly found in research applications. In order to derive these

estimates, four model selection criteria (i.e., the AIC, BIC, ABIC, and LRT) were

examined to determine their respective abilities to identify the correct number of latent

trajectory classes. This examination was conducted by comparing effective Type I error

rates to a nominal $\alpha = .05$. An additional goal of the study was to develop a means by

which estimates of power and accuracy could be generalized across multiple

experimental conditions and to conditions outside the scope of this study. Findings from

this study, with respect to the specific research questions posed earlier, are summarized

below.

1. *For a particular longitudinal design, how do standardized between-class differences in intercept and slope parameters (i.e., ds = 0.30, 0.50, 0.80, 1.00, and 1.20) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?*

Results from varying the standardized differences in intercept and slope

parameters between latent trajectory classes were consistent with previous research that

also examined this condition (Lo et al., 2001; Mendell et al., 1991, 1993). That is,

increased values of *d* results in positive increases with Type I error rate-adjusted

estimates of statistical power and accuracy. Results from this study indicate that, across

the examined range of values, the relationship is approximately linear, with adjusted

power increasing by .11 for every 0.10 increase in d (see Figure 7). However, for the

particular combination of experimental conditions used in this analysis (i.e., $N = 600$, six

equally spaced time points, no missing data, zero correlation between intercept and slope

parameters, and 60% of total variance explained by intercept and slope growth

parameters), growth parameters between classes have to be at least one standard

deviation apart to achieve .80 power. Therefore, researchers should consider more

favorable design conditions that would aid in the detection of latent trajectory classes

(e.g, a sample size greater than $N = 600$ or more than six time points) when classes are

hypothesized to be less than $d = 1.0$.

2. *For a particular longitudinal design, how does the percentage of variance explained by intercept and slope parameters (i.e., 40%, 50%, 60%, 70%, and 80%) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?*

As noted before, increasing the percentage of variance explained by intercept and

slope parameters is equivalent to reducing error variance, which results in positive, linear

increases in adjusted power and accuracy (See Figure 8). Similar to standardized

between-class differences in growth parameters, increasing the percentage of variance

explained by growth parameters by 10% (over the range of values examined here)

resulted in a corresponding .12 increase in adjusted power. Under the particular

combination of experimental conditions used here (i.e., $N = 600$, six equally spaced time

points, no missing data, zero correlation between intercept and slope parameters, and $d = 1.0$), adjusted power reaches .80 when the percentage of explained variance is 60%.

Consequently, percentages smaller than this require larger values for other design conditions to maintain adequate power.

3. *For a particular longitudinal design, how does the within-class correlation between intercept and slope parameters (rs = -.60, -.30, .00, .30, .60) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?*

The experimental conditions examined in this study assumed that intercept and slope growth parameters were orthogonal, that is that their correlation with each other was zero ($r = .00$). Relaxing this assumption resulted in only minor changes in adjusted power and accuracy when the correlation between growth parameters was between -.30 and +.60 (see Figure 9). For example, adjusted power increased from .046 at $r = .60$ to .180 at $r = -.30$. (It should be noted that estimates of adjusted power and accuracy for this experimental condition were derived from models requiring an additional degree of freedom [df = 14] compared to models for other experimental conditions, which fixed the covariance between intercepts and slopes at zero [df = 13]. Therefore, estimates of adjusted power and accuracy for this experimental condition are not directly comparable to estimates of power and accuracy derived for the other experimental conditions.)

However, changing the correlation from -.30 to -.60 had a relatively large impact on corresponding estimates of power and accuracy, increasing from .180 to .732, respectively. This asymmetric increase results from improved distinction between latent trajectory classes when correlations between intercepts and slopes are strongly negative. Given the particular population parameters for Class 1 (i.e., $\eta_{0k1} = 100$ and $\eta_{1k2} = 0$) and Class 2 (i.e., $\eta_{0k2} = 92.45$ and $\eta_{1k2} = -1.73$) in the "true" two-class (K2$_T$) model, larger negative within-class correlations between intercepts and slopes indicate that as

individual intercept values increase, their corresponding slope values decrease (and visa versa). The net result is greater separation between latent trajectory classes. By contrast, larger positive within-class correlations suggest that as individual intercept values increase (decrease), their corresponding slope values also increase (decrease), making latent trajectory class identification more difficult. These relationships are shown graphically in Figure 14 where variability in individual growth trajectories for Class 2, arising from the negative correlation between the growth parameters (shown in the top panel), leads to a greater overall distinction from Class 1 (here, held constant for illustration) than the variability that arises from the positive correlation between Class 2 growth parameters (shown in the bottom panel). It should be noted that the relationships depicted in Figure 14 are specific to the formation of latent trajectory classes used in this study. That is, other relationships between growth parameters could produce different results. However, the implication of this finding is that applied researchers should consider the degree and direction of covariation between growth parameters, as well as their mean values and unique variances, in their estimation of power and accuracy for latent trajectory class enumeration. This consideration is tempered by that fact that the correlation between growth parameters has to be strong to have substantial impact.

4. *For a particular population of individuals, how does sample size (i.e., $Ns = 200$, 400, 600, and 800), relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?*

Sample size is the characteristic most often associated with statistical power and typically the primary consideration that researchers face in creating research designs. Given this, sample size was analyzed with respect with three different between-class differences in growth parameters, namely $ds = 0.5$, 0.8, and 1.0. When $d = 0.5$, increasing

sample size from 400 to 1,000 participants had no discernable effect on adjusted power or accuracy (assuming six repeated measures, 60% of variance explained by growth parameters, no missing data, and zero correlation between intercept and slope parameters; see Figure 10). For this case, however, it should be noted that as sample size increases, both power and accuracy are theorized to eventually increase to their ultimate asymptotic limit. When $d = 0.8$, the result was a linear, positive relationship between sample size and power/accuracy, with power increasing by approximately .12 for every additional 200 participants. Furthermore, when $d = 1.0$, the corresponding increases in power and accuracy were nonlinear, increasing by approximately .25 for every additional 200 participants from Ns $= 200$ to 600 and by only .06 for every additional 200 participants from Ns $= 600$ to 1,000. These results suggest that examination of the relationship between a given experimental condition and power/accuracy as a mere two-dimensional phenomenon is insufficient and requires consideration of the interaction among all variables affecting power and accuracy.

5. *For a particular population of individuals, how does the number of repeated measures (i.e., 3, 4, 5, 6, 9, and 12 time points) relate to the statistical power and accuracy of the GMM to correctly enumerate a two-class latent trajectory solution?*

Another design condition that researchers often have to contend with is the number of repeated measures. Results for this experimental condition indicate that a substantial gain in power can be obtained by increasing the number of time points. For example, for the particular combination of experimental conditions examined here (i.e., $N = 600$, 60% of variance explained by growth parameters, no missing data, zero correlation between intercept and slope parameters, and $d = 1.0$), doubling the number of

time points from 3 to 6 resulted in almost a tenfold increase in adjusted power (i.e., from

.084 to .824). Although under a different set of conditions (e.g., smaller value of d, or

smaller sample sizes) the increase in adjusted power and accuracy may not be as

precipitous, increasing the number of time points (in the lower range of possible time

points) effectively results in improved detection of latent trajectory classes and,

consequently, improved power and accuracy. Again, the implication for applied

researchers is that less stringent design and population conditions (e.g., sample sizes less

than $N = 600$ or standardized between-class differences in growth parameters less than

1.0) are necessary to obtain adequate power (.80) when the number of repeated measures

is greater than six.

6. *For a particular population of individuals, how does planned missing data (i.e.,*
   25%, 50%, 75%, and 100% of cases missing either (a) every other time point or
   (b) every second and third time point, relate to the statistical power and accuracy
   of the GMM to correctly enumerate a two-class latent trajectory solution?

Incorporation of planned missingness into longitudinal research designs has been

shown to be an efficient way of examining a particular outcome over time (Graham,

Taylor, & Cumsille, 2001). Two such planned missingness designs were examined in this

study. In the first design, missing data were incorporated at every other time point; in the

second design missing data were incorporated at every 2nd and 3rd time point (see Table

5). In both designs, the total amount of missing data was varied in increments of 25%

from 0% (i.e., no missing data) to 100% (complete missing data for that time point).

Planned missingness was examined assuming $N = 600$ participants, 60% of variance

explained by growth parameters, zero correlation between intercept and slope parameters,

$d = 1.0$, and 12 repeated measures. In general, incorporation of planned missingness in

the data resulted in only small losses in adjusted power and accuracy. For the first design, increasing the amount of missing data from 0% to 75% resulted in no significant difference in estimates of adjusted power (.954 and .934, respectively). When 100% of cases were missing data for every other time point (equivalent to a six repeated-measure design with no missing data), adjusted power was still adequate (.824). This represents a 14% loss of power for a 50% saving of resources. For the second design, the corresponding losses in adjusted power and accuracy were greater; however, the estimates were still adequate when 75% of cases were missing data for every 2nd and 3rd time point (.861 and .802 for power and accuracy, respectively). This represents only a 10% loss in power for a 50% saving of resources. However, when the amount of missingness was increased to 100% (i.e., equivalent to a four repeated-measure design), the resulting decrement in power was substantial (i.e., adjusted power = .334). These findings suggest that, under the assumptions and specific population and design characteristics used in this study, latent trajectory class enumeration is robust to planned missing data when the overall number of repeated measures is at least six. This implies that researchers who incorporate planned missingness in their longitudinal designs can identify a given number of latent trajectory classes with considerably less resources (e.g., less repeated measurement occasions during the same time period) or potentially can identify a greater number of latent trajectory classes for the same amount of resources. Consequently, the use of planned missingness in longitudinal designs for GMMs is strongly recommended.

7. *Which model selection criterion (among the AIC, BIC, ABIC, and LRT) possesses the largest effective Type I error rates, and under which design and population conditions are Type I error rates the largest?*

In order to be comparable across the four examined model selection criteria, estimates of statistical power and accuracy were derived using a common (i.e., nominal) .05 Type I error rate. To accomplish this, effective Type I error rates for each criterion first were obtained via Monte Carlo simulation and were defined as the proportion of 500 replications that incorrectly rejected (a) a true null hypothesis of a one-class model estimated from a one-class pseudo-population ($K1_T$) in favor of a false alternative hypothesis of a two-class latent trajectory model estimated from a one-class pseudo-population ($K2_F$) and (b) a true null hypothesis of a two-class model estimated from a two-class pseudo-population ($K2_T$) in favor of an alternative false hypothesis of a three-class latent trajectory model estimated from a two-class pseudo-population ($K3_F$). Additionally, comparisons of log likelihood values for each of these model comparisons were used to identify 95% critical values used in obtaining the nominal .05 Type I error rates.

Effective Type I error rates for the $K1_T$ versus $K2_F$ model comparison are presented in Table 8 by experimental condition and quartiles of λ^2. Effective Type I error rates for the $K2_T$ versus $K3_F$ model comparison similarly are presented in Table 9. For both sets of comparisons, the BIC was shown to be extremely conservative, with observed Type I error rates of approximately .001; that is, under repeated samples from the same population, the BIC would select the erroneous $K2_F$ or $K3_F$ model only one time in a thousand. However, the price that is paid for such veracity is correspondingly lower estimates of nominal power and accuracy. For example, for Population Y (described in

Table 4), adjusted power (i.e., at a nominal .05 Type I error rate) was .954. The

corresponding estimate of power at an effective .001 Type I error rate was .670.

In contrast, the AIC was observed to have very liberal effective Type I error rates,

ranging from .147 for $K1_T$ versus $K2_F$ (averaged across different sample sizes) to .244 for

$K2_T$ versus $K3_F$ (averaged across various correlations between intercepts and slopes).

Consistent with previous research (Lin & Dayton, 1997; Masyn & Brown, 2001; Yang,

1999), these findings indicate a tendency for the AIC to favor models with larger

numbers of latent trajectory classes. This tendency is demonstrated by a comparison

between power estimated at effective and nominal Type I error rates, which showed that

effective power for the AIC was, on average, 14% greater than power adjusted at the

nominal $\alpha = .05$. Although increased power is generally desirable, for the AIC, power is

obtained at the price of overly rejecting more parsimonious models.

Effective Type I error rates for the LRT were only slightly larger than the nominal

$\alpha = .05$. For example, effective Type I error rates for the LRT ranged from .070 for $K1_T$

versus $K2_F$ (averaged across different sample sizes at $d = 0.80$) to .141 for $K2_T$ versus

$K3_F$ (averaged across various correlations between intercepts and slopes). This finding

holds promise for the recently developed adjusted-LRT (Lo et al., 2001), which has just

begun to be examined via Monte Carlo simulation (Masyn & Muthén, 2002).

The ABIC consistently demonstrated observed Type I error rates that were closest to

the nominal $\alpha = .05$. For example, across all experimental conditions, the average

effective Type I error rates for the ABIC were .046 for $K1_T$ versus $K2_F$ and .062 for $K2_T$

versus $K3_F$. Effective Type I error rates for the ABIC ranged from .032 for $K1_T$ versus

$K2_F$ (averaged across different sample sizes at $d = 0.80$) to .077 for $K2_T$ versus $K3_F$

(averaged across different sample sizes at $d = 1.0$). Consequently, among the four model selection criteria examined in this study, the ABIC is recommended as the model selection criterion that provides the best balance between nominal Type I and Type II errors in enumerating latent trajectory classes.

8. *How do statistical power and accuracy relate to the overall separation between one- and two-latent trajectory class models, as measured by λ^2?*

A measure of the overall separation between latent trajectory classes (λ^2) was operationalized as -2 times the difference in log likelihoods between the competing $K1_F$ and $K2_T$ models, averaged across the 500 replications, minus df $= 3$. Resulting estimates of power and accuracy (separately) were regressed on λ^2 to yield functions by which power and accuracy, respectively, can be estimated for all values of λ^2. That is, the functional relationship between power and accuracy and λ^2 derived in this study provides a means to determine how "far apart" latent trajectory classes have to be in order for researchers to accurately and reliably detect them. Thus, λ^2 acts as a latent trajectory class analog to other measures of multivariate separation or distinction between group centroids (e.g., Mahalanobis distance or Kullback-Leibler discrepancy functions). More specifically, λ^2 acts as an approximation to a noncentrality parameter based on a chi-square distribution, in line with power estimation for traditional latent variable models (MacCallum, Browne, & Sugawara, 1996; Saris & Satorra, 1993; Satorra & Saris, 1985). In practical terms, any change in design or population characteristics that increases the ability to discern latent trajectory classes is reflected in a greater difference in log likelihoods between "true" and "false" models, resulting in larger values of λ^2, and consequently, greater power and accuracy.

71

The derived functional relationships between estimates of adjusted power and accuracy and λ^2 are described by positive, monotonic functions with significant negative quadratic terms representing diminishing returns to power and accuracy as values of λ^2 increase (see Table 11 and Figure 13). Across all 36 experimental conditions examined in this study, mean $\lambda^2 = 9.54$ ($SD = 7.14$, range = 0.75 to 22.38). Both functions demonstrated a high degree of fit to the data (R^2s = .98 and .99, for adjusted power and accuracy, respectively), resulting in relatively small confidence intervals for predicted values.

Alternatively, the empirically derived estimates of statistical power can be compared to power estimates derived from a noncentral chi-square distribution (as shown in Table 11). Although power estimates from the Monte Carlo simulations are somewhat smaller than their asymptotically correct counterparts, they provide a good approximation to the distribution, and consequently, demonstrate the potential of the noncentral chi-square distribution as a method for determining power for latent trajectory class enumeration. That is, researchers can use estimates of λ^2 as an alterative to more extensive Monte Carlo simulations or the aforementioned mathematical functions generated from the simulations conducted in this study. For example, for a $\lambda^2 = 10.91$, α = .05, and df = 3, power based on a noncentral chi-square distribution is .80 (derived from either tables of the distribution [e.g., Haynam, Govindarajulu, & Leone, 1973] or computer software [e.g., NCSS statistical software]). By comparison, for the same λ^2, α, and df, adjusted power based on the Monte Carlo simulations from this study is estimated at .77. These relationships are illustrated in examples of two hypothetical research applications and in the analysis of real data from the motivating example.

Application of Results

      Scenario #1. Assume that a researcher has conducted a GMM of reading achievement trajectories from annual assessment of 125 students during 1st through 4th grades. Further assume that results of the analysis, based on penalized information criteria, support a two latent trajectory class model with Class 1 representing students whose grade-equivalency scores remain relatively constant over time (i.e., a no-change class) and Class 2 representing students whose scores decrease over time (i.e., a significant negative linear slope class). From theory, the researcher hypothesizes that there may be an additional third class of students, those whose reading achievement scores increase over time; however, this class would represent a smaller proportion of the total number of students and, given the relatively small sample size used in the analysis, there may not have been sufficient power to detect this class. The researcher decides to conduct a power analysis using Monte Carlo simulation to determine if this was, in fact, the case.

      First, the researcher specifies the null hypothesis as a "false" two-class solution derived from a "true" three-class pseudo-population and translates hypothesized values for population and design characteristics into the appropriate computer input statements. For example, the researcher may assume the following: that the underlying population distributions for the outcome reading achievement scores is multivariate normal; mixing proportions of 60% for Class 1, 30% for Class 2, and 10% for the hypothesized Class 3 (i.e., students with increasing scores); that 50% of the total variance consists of variation in the Level-2 growth parameters; that intercept and slope growth parameters are correlated at -.10; that between-class intercepts and slopes are 1.0 standard deviation unit

apart; no missing data; and design characteristics consisting of 125 students and four equally-spaced repeated measurement occasions. Next, the researcher specifies the alternative hypothesis as a "true" three-class solution derived from the same "true" three-class pseudo-population as for the null hypothesis (accomplished by the use of a common random number seed in generating the simulated sample data). The same population and design characteristics are similarly incorporated into the corresponding computer input statements. Furthermore, the researcher decides that 500 replications will provide a sufficient confidence interval for estimated power.

At this point, the researcher has several options. First, he/she can link resultant output files by replication, select the optimal model via an a priori decision rule, and calculate a brute-force estimate of power as the proportion of replications that correctly identify the true alternative hypothesis. Power estimates constructed in this way benefit from not having to rely on a particular distribution representing the discrepancy between competing models, however, they may be sensitive to fluctuations in the process of generating the random samples. Alternatively, the researcher can estimate power by comparing values of λ^2 calculated as -2 times the average difference in log-likelihoods between models representing null and alternative hypotheses (minus df). Values of λ^2 based on a noncentral chi-square distribution provide an asymptotic upper bound on estimates of power. Values of λ^2 based the mathematical function generated from the Monte Carlo simulations used in this study provide a lower bound on power estimates given the restrictions imposed by finite population and design characteristics. For the above scenario, assume that log likelihood values of -1787.91 and -1785.65 are generated from models representing the null and alternative hypotheses, respectively. Consequently,

74

a value of 1.52 is calculated for λ^2. Relating this value to the mathematical functions generated in this study yields an estimated power of .080. By comparison, from the noncentral chi-square distribution with a Type I error rate of .05 and df = 3, power is estimated to be .155. The researcher concludes, therefore, that the failure to detect the hypothesized third class of students may be due to insufficient power.

     Scenario #2. Based on information from Scenario #1, assume that the researcher now plans a new study with the goal of identifying a latent trajectory class of students with increasing reading achievement scores (i.e., Class 3). Although the researcher still is limited to four annual repeated measures, as well as the other population and design constraints mentioned above, he/she can increase the sample size. Therefore, a new power analysis is conducted to determine how many participants need to be sampled to achieve power of .80. One alternative is to conduct additional Monte Carlo simulations with the researcher incrementally increasing sample size until the desired level of power is obtained. However, a less labor intensive procedure is to rely on either the functional relationship between adjusted power and λ^2 derived from this study or a noncentral chi-square distribution. For example, from the Monte Carlo-based mathematical function, adjusted power of .80 corresponds to a λ^2 value of 12.35. From the noncentral chi-square distribution with a Type I error rate of .05 and df = 3, power of .80 corresponds to a λ^2 value of 10.91. As the relationship between sample size (*N*) and λ^2 is linear, the value of N/λ^2 (i.e., 125/1.52 = 82.24 in this scenario) can be multiplied by values of λ^2 (12.35 and 10.91, respectively) that correspond to the desired level of power (.80) to provide estimates of the required sample size (*N*s = 1,016 and 897, respectively). Therefore, given the particular population and design characteristics assumed in this scenario, the

researcher will have to sample between 897 and 1,016 students to have sufficient power to enumerate a third latent trajectory class. The difference between these two estimates reflects the difference between a "best case scenario" of estimated λ^2 following perfectly the theorized noncentral chi-square distribution versus the less-than-perfect limitations of its finite sampling distribution.

Findings from the Motivating Example

This study included a motivating example that relied on a specific set of design and population characteristics from an evaluation of a *system-of-care* demonstration project to illustrate the process of estimating power using Monte Carlo simulation. In turn, these estimates were compared to estimates of power predicted from (a) the functional relationship between λ^2 and power derived from the 36 experimental conditions examined earlier and (b) a noncentral chi-square distribution. Although the functional relationship is based on a comparison of a "true" two-class model compared to a "false" one-class model, this relationship is extended to also test the power to detect a "true" three-class model over a "false" two-class model.

Monte Carlo simulation of data patterned after the demonstration project evaluation resulted in a high degree of power to detect two- and three-class latent trajectory models (i.e., adjusted power = .894 and .914, respectively). Comparison of "true" models with their respective "false" models resulted in λ^2 values of 18.88 and 34.49, respectively. Inserting these values into the functional relationship produced predicted power estimates of .957 and 1.00, respectively. These estimates were somewhat higher than estimates produced from brute-force Monte Carlo simulation, however, the predicted estimate for the two-class model was within the 95% confidence band for the

76

function (i.e., .874 to 1.00 for $\lambda^2 = 18.88$). The predicted estimate for the three-class

model, however, was just outside its respective 95% confidence band (i.e., .917 to 1.00

for $\lambda^2 = 34.49$). By comparision, based on the noncentral chi-square distribution with a

Type I error rate of .05 and df = 3, power was estimated to be .967 for the two-class

model and 1.00 for the three-class model. These results imply a successful test of the

functional relationship between λ^2 and power for the two-class case and a marginally

successful test for the three-class case. This is not surprising given that the function was

derived from examination of "true" two class models to "false" one-class models.

Consequently, these results imply that the generalizability of this function may be in

question for predictions of "true" three-class models over "false" two-class models. This

latter comparison, as well as for comparisons of higher numbers of latent trajectory

classes with their k_0 -1 counterparts may require derivation of additional, comparison-

specific, functional relationships.

Limitations and Directions for Future Research

The experimental conditions used in this study were chosen to represent variables

that an applied researcher would have to address in estimating power and accuracy

among GMMs with varying numbers of latent trajectory classes. Nonetheless, limitations

to this study still exist. First, the study did not examine growth trajectories and variance-

covariance structures not explicitly covered by the two-class population distributions

used in this study. For example, trajectories that cross over in time, quadratic, cubic, or

other nonlinear trajectories, trajectories derived from nonnormal distributions, and

noninvariant residual variances within and between latent trajectory classes were not

77

examined. Estimation of power and accuracy for these types of models is left as an open question to be examined in future research.

Second, the current study did not examine the role of unequal mixing proportions on statistical power or accuracy. That is, all experimental conditions examined in this study assumed a 50-50 split of cases between latent trajectory classes. Although previous research (Lo et al., 2001; Masyn, 2001; Mendell et al., 1991, 1993) has shown little relationship between various combinations of mixing proportions and the ability to identify latent classes, only the Masyn study examined different mixing proportions in the context of the GMM, and that study was conducted within a restricted set of experimental conditions. Moreover, practical applications of the GMM seldom, if ever, result in evenly divided mixing proportions; thus, the role varying mixing proportions in the GMM remains as another avenue for future research.

Third, the relationship between model covariates and estimates of statistical power was not explored in this study. Although other studies have shown covariates to be related to both power (Allison et al., 1997; Jo, 2002) and design efficiency (Raudenbush, 1977), the relationship between model covariates and latent trajectory class enumeration also remains as an area to be explored.

Fourth, as maximum likelihood solutions generated by the GMM may be sensitive to different starting values, this study also was limited by lack of variation in starting values used for each of the tested models. To address this, a random sample of nonconvergent replications for the three-class model was reanalyzed with a second, more disparate, set of starting values. Results indicated that there was no greater likelihood of model convergence with the second set of starting values than with the first set of starting

values. This test, however, did not take into consideration that the applied researcher typically has more information and flexibility in choosing starting values than the uniform sets of values used in this study.

Fifth, this study did not examine combinations of all values among the six selected experimental conditions; to do so would have constituted 22,500 unique combinations of experimental conditions, requiring an impracticable and inefficient allocation of resources. However, this study did examine the interaction between levels of between-class differences in growth parameters and sample size as they related to estimates of power and accuracy (shown in Figure 10). This limitation also was addressed by explicating the relationship between the separation and distinction between latent trajectory classes (i.e., λ^2) and corresponding estimates of power and accuracy from which new estimates can be derived for particular combinations of experimental conditions not explicitly examined in this study.

Sixth, the role of nesting or clustering of individuals into groups, which commonly exists in developmental and intervention studies (e.g., sampling participants from classroom, schools, or communities), was not examined in this study. Although it is possible that these "Level-3" effects could potentially impact latent trajectory class enumeration through the partitioning the amount of total variation into within- and between-group components, current methods to conduct growth mixture modeling within a hierarchical/multilevel framework are not available.

Finally, the procedure used in this study for generating power and accuracy estimates relied on effective Type I error rates ($\alpha = .05$) generated via "brute force" Monte Carlo simulation. In practice, knowledge of Type I error rates derived from a more

rigorous theoretical perspective would serve to greatly expedite the process of estimating power and accuracy. For example, strong distributional theory for the LRT in non-nested models, such as that developed by Vuong (1989) and incorporated in the adjusted Lo, Mendell, and Rubin-LRT (LMR-LRT, 2001), has been employed to develop new tests of fit for models based on mixture distributions. Furthermore, the LMR-LRT has been implemented in recent versions of statistical software (i.e., Mplus v.2.1; Muthén & Muthén, 2001) and has been shown to be an effective alternative to existing model selection criteria (Masyn & Muthén, 2002).

Conclusions

Several general conclusions are gleaned from this study. First, results of the Monte Carlo simulations to determine effective Type I error rates demonstrate that enumeration of latent trajectory classes is strongly influenced by the choice of model selection criteria. Consistent with the recommendations of Yang (1999), the use of the ABIC is advocated over the other three criteria examined in this study. Additionally, it is noted that the LRT holds promise as a parametric alternative to penalized information criteria. Moreover, use of the BIC and AIC, given their respective conservative and liberal Type I error rates, can be used to establish lower and upper bounds on the number of latent trajectory classes in longitudinal data.

Second, results of the Monte Carlo simulations for each experimental condition indicate that, as hypothesized, the ability to correctly enumerate latent trajectory classes is positively related to greater (a) differences in intercept and slope parameters between classes, (b) percentages of variance explained by intercept and slope parameters, (c) sample sizes, and (d) numbers of repeated measures. Among these four conditions,

increasing the number of repeated measures appears to give the greatest increase in adjusted power and accuracy. However, results from the analysis of sample size indicate that statistical power and accuracy need to be estimated in the context of multiple conditions. Consequently, sample size "rules of thumb" for the GGMM, such as 100 cases per latent trajectory class (Lawrence & Wagstaff, 2003), may be misleading for latent trajectory class enumeration.

Third, results of the Monte Carlo simulations across the 36 combinations of experimental conditions examined in this study suggest that the conditions necessary for adequate power and accuracy (adjusted at a nominal $\alpha = .05$) are relatively expensive, in terms of the data requirements necessary to reliably enumerate latent trajectory classes. For example, to achieve an adjusted power of .80 requires a sample size of 600 participants, six repeated measures, 60% of variation attributed to intercept and slope growth parameters, a between-class difference of 1.0 standard deviation for growth parameters, and no missing data. Less information on any one of these conditions would require more information from other conditions in order to maintain adequate power and accuracy. Nonetheless, from a review of research applications of the GMM, it also is noted that researchers have been successful in finding mixture distributions in longitudinal data. This success may be due, in part, to the partitioning of variability in growth parameters that arises as latent trajectory classes are identified. Consequently, relatively large (i.e., > 1.0) standardized between-class differences in growth parameters should not be surprising in such research applications.

Finally, and most importantly, examination of λ^2 derived from the difference in log likelihoods between the "false" one-class and "true" two class models demonstrates

81

the utility of relating power and accuracy to a measure of the degree of between-class separation (e.g., as a chi-square based noncentrality parameter). This conclusion has implications for researchers wishing to derive estimates of statistical power or accuracy in that the process of obtaining such estimates can be greatly simplified; summary log likelihood values can be compared without the necessity of linking information criteria across competing models (i.e., models with differing numbers of latent trajectory classes) and identifying the optimal model for each replication. When applied to a motivating example using "real" design and population characteristics, λ^2 proved successful in predicting actual levels of power to detect a two-class mixture distribution over a one-class distribution. However, the generalizability of this measure did not fully extend to the ability to detect a three-class mixture distribution over a two-class distribution. Therefore, it is preferable that estimates of λ^2 be developed from the specific model comparisons for which estimates of power and accuracy are intended to be predictive.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, VT: University of Vermont Department of Psychiatry.

Akaike, H. (1973). Statistical predictor identification. *Annals of the Institute Statistical Mathematics*, *22*, 203-217.

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2*, 20-33.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics 29,* 451-457.

Biernacki, C., Celeux, G., & Govaert, G. (1998). Assessing a mixture model for clusterings with the integrated classification likelihood. *Technical Report No. 3521*. Rhône-Alpes: INRIA.

Biernacki, C., Celeux, G., & Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters, 20*, 267-272.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345-370.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal, 13,* 195-212.

Chuang, R., & Mendell, N. R. (1997). The approximate null distribution of the likelihood ratio test for a mixture of two bivariate normal distributions with equal covariance. *Communications in Statistics—Simulation and Computation, 26*, 631-648.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences.* (pp. 311-359). New York: Plenum Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

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

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

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structure analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*, 512-519.

Curran, P. J., West, S. G., & Finch, J. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Dacunha-Castelle, D., & Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli, 3*, 279-299.

Dedrick, R. F., Greenbaum, P. E., Vaughn, D. A. (April, 2001). *Longitudinal analysis of reading achievement trajectories of low achieving second graders: General growth mixture modeling*. Presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Dennis, M. L., Lennox, R. D., & Foss, M. A. (1997). Practical power analysis for substance abuse health services research. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention*. Washington, DC: American Psychological Association.

Dumenci, L., & Windle, M. (2001). Cluster analysis as a method of recovering types of intraindividual growth trajectories: A Monte Carlo study. *Multivariate Behavioral Research, 36*, 501-522.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Everitt, B. S. (1988). A Monte Carlo investigation of the likelihood ratio test for the number of classes in latent class analysis. *Multivariate Behavioral Research, 23*, 531-538.

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

Furman, W. D. & Lindsay, B. G. (1994). Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics and Data Analysis, 17,* 473-492.

Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods and Research, 21,* 132-160.

Golden, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology, 44*, 153-170.

Goldstein, H. (1995). Hierarchical data modeling in the social sciences. *Journal of Educational and Behavioral Statistics, 20*, 201-224.

Graham, J. W., Taylor, B. J., Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change: Decade of behavior*. (pp. 335-353). Washington, DC: American Psychological Association.

Greenbaum, P. E., & Brown, E. C. (May, 2002). *Growth mixture modeling and propensity analysis: Evaluating an integrated services program for children with serious emotional disorders*. Presented at the 10th Annual Meeting of the Society for Prevention Research, Seattle, WA.

Greenbaum, P. E., Dedrick, R. F., & Brown, E. C. (April, 2000a). *Group differences in developmental trajectories of externalizing behaviors for children with emotional disturbance: Growth mixture modeling*. Presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Greenbaum, P. E., Dedrick, R. F., & Brown, E. C. (August, 2000b). *Growth mixture modeling of delinquent behavior developmental trajectories*. Presented at the annual meeting of the American Psychological Association, Washington, DC.

Greenbaum, P. E., Del Boca, F. K., Darkes, J., and Goldman, M. S. (2001). Unpublished manuscript. University of South Florida.

Gutierrez, R. G., Carroll, R. J., Wang, N., Lee, G. H., & Taylor, B. H. (1995). Analysis of tomato root initiation using a normal mixture distribution. *Biometrics, 51,* 1461-1468.

Jo, B. (2002). Statistical power in randomized intervention studies with noncompliance. *Psychological Methods, 7*, 178-193.

Haughton, D. M. A., Oud, J. H. L., & Jansen, R. A. R. G. (1997). Information and other criteria in structural equation model selection. *Journal of Statistical Computation and Simulation, 26,* 1477-1516.

Haynam, G. E., Govindarajulu, Z., & Leone, F. C. (1973). Tables of the cumulative noncentral chi-square distribution. In H. L. Harter and D. B . Owen (Eds.), *Selected tables in mathematical statistics* (Vol. 1, pp. 1-78). Providence, RI: Mathematical Statistical Society.

Heckman, J. J., Robb, R., & Walker, J. R. (1990). Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the method of moments. *Journal of the American Statistical Association, 85,* 582-589.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Ichikawa, M., & Konishi, S. (1999). Model evaluation and information criteria in covariance structure analysis. *British Journal of Mathematical & Statistical Psychology, 52*, 285-302.

Kellam, S. G. & Rebok, G. W. (1992). Building developmental and etiological theory through epidemiologically based preventive intervention trials. In J. McCord & R. E. Tremblay (Eds.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp.162-195). New York: Guilford Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Lawrence, F. R. & Wagstaff, D. A. (April, 2003). *Sample size requirements for General Growth Mixture Models*. Presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Li, F., Duncan, T. E., Hops, H. (2001). Examining developmental trajectories in adolescent alcohol use using piecewise growth mixture modeling analysis. *Journal of Studies on Alcohol, 62*, 199-210.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22*, 249-264.

Lindsay, B. G., & Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association, 87*, 785-794.

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

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 2*, 130-149.

Loehlin, J. C. (1998). *Latent variable models* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Masyn, K. (September, 2001). *Further simulation studies of latent class enumeration in the GGMM framework.* Presented at the annual meeting of the Prevention Science Methodology Group. University of California Los Angeles.

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

Masyn K., & Muthén, B. (November, 2000). *Simulation procedures in Mplus*. Presented at the weekly teleconference meeting of the Prevention Science Methodology Group.

Masyn K., & Muthén, B. (April, 2002). *An examination of the Lo, Mendel, and Rubin likelihood ratio test using Monte Carlo simulation in Mplus*. Presented to the Annual Meeting of the American Educational Research Association, Seattle, WA.

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

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

McLachlan, G. J., & Peel, D. (1997). On a resampling approach to choosing the number of components in normal mixture models. In L. Billard and N. I. Fisher (Eds.), *Computing science and statistics, Vol. 28* (pp. 260-266). Fairfax Station, VA: Interface Foundation of North America.

McLachlan, G. J., & Peel, D. (2001). *Finite mixture models*. New York: John Wiley & Sons, Inc.

Mendell, N. R., Finch, S. J., & Thode, H. C. (1993). Where is the likelihood ratio test powerful for detecting two component normal mixtures? *Biometrics, 49*, 907-915.

Mendell, N. R., Thode, H. C., & Finch, S. J. (1991). The likelihood ratio test for the normal mixture problem: Power and sample size analysis. *Biometrics, 47*, 1143-1148.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Iowa City, IA: Sage University Press.

Muthén, B. O. (2001a). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-22). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Muthén, B. O. (2002). *Using Mplus Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models*. Mplus Web Notes Number 1 [On-line]. Available: www.statmodel.com.

Muthén, B. O., & Asparouhov, T. (2002). *Using Mplus Monte Carlo simulations in practice: A note on non-normal missing data in latent variable models*. Mplus Web Notes Number 2 [On-line]. Available: www.statmodel.com.

Muthén, B. O., Brown, C. H., Masyn, K., Jo, B., Khoo, S., Yang C., Wang, C., Kellam, S. G., Carlin, J. B., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 9*, 456-678.

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

Muthén, B. O., and Khoo, S. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual Differences, 10*, 73-101.

Muthén, B. O., Khoo, S. T., Francis, D., & Kim Boscardin, C. (2000). Analysis of reading skills development from Kindergarten through first grade: An application of growth mixture modeling to sequential processes. In S. R. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications*. Mahaw, NJ: Lawrence Erlbaum Associates.

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

Muthén, L. K., & Muthén, B. O. (2001). Mplus Users' Guide (Second edition). [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*, 599-620.

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

NCSS Statistical Software (1995). [shareware: www.ncss.com] Kaysville, Utah.
Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric group-based approach. *Psychological Methods, 2*, 139-157.

ORC Macro International (1995). *Comprehensive Community Mental Health Services for Children and Their Families Program.* (W. Holden, Principal Investigator). Atlanta, GA.

Olmos, A., & Hutchinson, S. R. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling, 5*, 344-364.

Oxford, M. L., Gilchrist, L. D., Morrison, D. M., Gillmore, M. R., Lohr, M. J., & Lewis, S. M. (2003). Alcohol use among adolescent mothers: Heterogeneity in growth curves, predictors, and outcomes of alcohol use over time. *Prevention Science, 4*, 15-26.

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

Petras, H., Furr-Holden, C. D., Clark, M. D., Brown, C. H., Ialongo, N., & Kellam, S. (May, 2001). Aggressive/disruptive behavior in school children and school removal: An application of growth mixture modeling. Presented at the 9th annual meeting of the Society for Prevention Research, Washington, DC.

Raudenbush, S. W. (1977). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications, Inc.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14,* 465-471.

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

SAS Institute Inc. (1988). *SAS/IML User's Guide, Release 6.03 Edition*. Cary, NC: SAS Institute Inc.

SPSS Inc. (2000). SPSS for Windows 10.0 [Computer software]. Chicago, IL: SPSS Inc.

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181-204). Newbury Park, CA: Sage.

Satorra, A., & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 51, 83-90.

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

Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35*, 137-167.

Stroul, B. A., & Friedman, R. M. (1986). *A system of care for children and youth with severe emotional disturbances*. (Revised edition). Washington, DC: Georgetown University Child Development Center, CASSP Technical Assistance Center.

Tanguma, J. (2001). Effects of sample size on the power of selected fit indices: A graphical approach. *Educational and Psychological Measurement, 61*, 759-776.

Thode, H. C., Finch, S. J., & Mendell, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio for a mixture of two normals. *Biometrics, 44*, 1195-1201.

Titterington, D. M., Snith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, U.K.: John Wiley & Sons.

Tucker, J. S., Orlando, M., & Ellickson, P. L. (2003). Patterns and correlates of binge drinking trajectories from early adolescence to young adulthood. *Health Psychology, 22*, 79-87.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57*, 307-333.

Wang, C. P., Brown, C. H., Carlin, J., & Dagne, G. (2001). *Model diagnostics for general growth mixture models: An examination of differential trajectories of aggressive behavior in children.* Unpublished manuscript.

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.

Wolfe, J. H. (1971). A Monte Carlo study of sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Technical Bulletin STB 72-2*. San Diego: U.S. Naval Personnel and Training Research Laboratory.

Xitao, F., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56-83.

Xitao, F., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational & Psychological Measurement, 58,* 701-735.

Yang, C. C. (1999). Finite Mixture Model Selection with Psychometric Applications. (Doctoral dissertation, University of California, Los Angeles, 1999). *Dissertation Abstracts International, 59(9-A),* 3421.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44*, 41-61.

Appendices

Mplus Input Syntax for Analysis of Growth Mixture Model for Population Y

One-class GMM estimated from a one-class pseudo-population (K1$_T$)

```
TITLE:  Monte Carlo simulation of GMM for condition hba
        -- specifies a true one-class model tested against
        a true one-class pseudo-population


MONTECARLO: NAMES ARE score0-score11;
            NOBSERVATIONS = 600;
            NREPS = 500;
            SEED = 87809435;
            NCLASSES=1;
            GCLASSES=1;
            SAVE = c:\mplus3\hba3pgms\hbaREPKX.DAT;
            RESULTS = c:\mplus3\hba3pgms\hbaOUTKX.DAT;

ANALYSIS:   TYPE = MIXTURE;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
                    score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt*57 slope*3;
            intercpt with slope@0;
            score0*40 score1*40 score2*40 score3*40 score4*40
            score5*40 score6*40 score7*40 score8*40 score9*40
            score10*40 score11*40;


        %c#1%
            [intercpt*96.225 slope*-0.865];



MODEL: %overall%
            intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
            score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt with slope@0;


        %c#1%
            [intercpt*96.225 slope*-0.865];

OUTPUT:     TECH9;
```

One-class GMM estimated from a two-class pseudo-population (K1$_F$)

```
TITLE:  Monte Carlo simulation of GMM for condition hba
        -- specifies a false one class model tested against
        a true two class pseudo-population


MONTECARLO: NAMES ARE score0-score11;
            NOBSERVATIONS = 600;
            NREPS = 500;
            SEED = 87809435;
            NCLASSES=1;
            GCLASSES=2;
            SAVE = c:\mplus3\hba3pgms\hbaREPK1.DAT;
            RESULTS = c:\mplus3\hba3pgms\hbaOUTK1.DAT;

ANALYSIS:   TYPE = MIXTURE;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
                    score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt*57 slope*3;
            intercpt with slope@0;
            score0*40 score1*40 score2*40 score3*40 score4*40
            score5*40 score6*40 score7*40 score8*40 score9*40
            score10*40 score11*40;

             %c#1%
            [intercpt*100 slope*0.00];

             %c#2%
            [intercpt*92.45 slope*-1.73];


MODEL: %overall%
             intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
            score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt with slope@0;

              %c#1%
              [intercpt*96.225 slope*-0.865];



OUTPUT:     TECH9;
```

Two-class GMM estimated from a two-class pseudo-population (K2$_T$)

```
TITLE:   Monte Carlo simulation of GMM for condition hba
         -- specifies a true two class model tested against
         a true two class pseudo-population

MONTECARLO: NAMES ARE score0-score11;
            NOBSERVATIONS = 600;
            NREPS = 500;
            SEED = 87809435;
            NCLASSES=2;
            GCLASSES=2;
            SAVE = c:\mplus3\hba3pgms\hbaREPK2.DAT;
            RESULTS = c:\mplus3\hba3pgms\hbaOUTK2.DAT;

ANALYSIS:   TYPE = MIXTURE;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
             intercpt BY score0-score11@1;
             slope BY score0@0 score1@1 score2@2 score3@3 score4@4
              score5@5
             score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
             [score0-score11@0];
             intercpt*57 slope*3;
             intercpt with slope@0;
             score0*40 score1*40 score2*40 score3*40 score4*40
             score5*40 score6*40 score7*40 score8*40 score9*40
             score10*40 score11*40;

              %c#1%
             [intercpt*100 slope*0.00];

              %c#2%
             [intercpt*92.45 slope*-1.73];

MODEL: %overall%
              intercpt BY score0-score11@1;
             slope BY score0@0 score1@1 score2@2 score3@3 score4@4
              score5@5
             score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
             [score0-score11@0];
             intercpt with slope@0;

              %c#1%
             [intercpt*100 slope*0.00];

              %c#2%
             [intercpt*92.45 slope*-1.73];

OUTPUT:     TECH9;
```

95

Two-class GMM estimated from a one-class pseudo-population (K2$_F$)

```
TITLE:  Monte Carlo simulation of GMM for condition hba
        -- specifies a false two class model tested against
        a true one class pseudo-population


MONTECARLO: NAMES ARE score0-score11;
            NOBSERVATIONS = 600;
            NREPS = 500;
            SEED = 87809435;
            NCLASSES=2;
            GCLASSES=1;
            SAVE = c:\mplus3\hba3pgms\hbaREPK0.DAT;
            RESULTS = c:\mplus3\hba3pgms\hbaOUTK0.DAT;

ANALYSIS:   TYPE = MIXTURE;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
                    score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt*57 slope*3;
            intercpt with slope@0;
            score0*40 score1*40 score2*40 score3*40 score4*40
            score5*40 score6*40 score7*40 score8*40 score9*40
            score10*40 score11*40;

             %c#1%
             [intercpt*96.225 slope*-0.865];

MODEL: %overall%
            intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
            score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt with slope@0;


             %c#1%
            [intercpt*100 slope*0.00];

             %c#2%
            [intercpt*92.45 slope*-1.73];


OUTPUT:     TECH9;
```

Three-class GMM estimated from a two-class pseudo-population (K3$_F$)

```
TITLE:   Monte Carlo simulation of GMM for condition hba
         -- specifies a false three class model tested against
         a true two class pseudo-population

MONTECARLO: NAMES ARE score0-score11;
            NOBSERVATIONS = 600;
            NREPS = 500;
            SEED = 87809435;
            NCLASSES=3;
            GCLASSES=2;
            SAVE = c:\mplus3\hba3pgms\hbaREPK3.DAT;
            RESULTS = c:\mplus3\hba3pgms\hbaOUTK3.DAT;

ANALYSIS:   TYPE = MIXTURE;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4
             score5@5
            score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt*57 slope*3;
            intercpt with slope@0;
            score0*40 score1*40 score2*40 score3*40 score4*40
            score5*40 score6*40 score7*40 score8*40 score9*40
            score10*40 score11*40;

             %c#1%
            [intercpt*100 slope*0.00];

             %c#2%
            [intercpt*92.45 slope*-1.73];

MODEL: %overall%
                intercpt BY score0-score11@1;
            slope BY score0@0 score1@1 score2@2 score3@3 score4@4 score5@5
            score6@6 score7@7 score8@8 score9@9 score10@10 score11@11;
            [score0-score11@0];
            intercpt with slope@0;

             %c#1%
            [intercpt*92.45 slope*-1.73];

             %c#2%
            [intercpt*96.225 slope*-0.865];

             %c#3%
            [intercpt*100.00 slope*0.00];

OUTPUT:     TECH9;
```

SAS Program Used to Create Simulated Data for Population Y

```
*********************************************************************;
* SAS program to generate a simmulated data set                    *;
* Condition #24X (2 classes)                                       *;
* class 1: ns = 300, nt = 12, intrcpt = 100.0,  slope=  0.00       *;
* class 2: ns = 300, nt = 12, intrcpt =  92.45, slope= -1.73       *;
*********************************************************************;

proc iml;
create m1 var{person time score wave};
do per = 1 to 600;
  if per <= 300 then do;
     time = {0,1,2,3,4,5,6,7,8,9,10,11};
     length = nrow(time);
     person = repeat(per,length);
     x = round(1000000*ranuni(0));
     resid = armasim(0,0,0,6.32,length,x);
     intrcept = repeat((normal(0)*7.55)+100.0,length);
     slope = (normal(0)*1.73) - 0.00;
     score = intrcept + slope*time + resid;
  end;
  else do;
     time = {0,1,2,3,4,5,6,7,8,9,10,11};
     length = nrow(time);
     person = repeat(per,length);
     x = round(1000000*ranuni(0));
     resid = armasim(0,0,0,6.32,length,x);
     intrcept = repeat((normal(0)*7.55)+92.45,length);
     slope = (normal(0)*1.73) + -1.73;
     score = intrcept + slope*time + resid;
  end;
append;
end;
close m1;
quit;

data d1;
  set work.m1;
  score11 = score;
  score10 = lag(score);
  score9 = lag2(score);
  score8 = lag3(score);
  score7 = lag4(score);
  score6 = lag5(score);
  score5 = lag6(score);
  score4 = lag7(score);
  score3 = lag8(score);
  score2 = lag9(score);
  score1 = lag10(score);
  score0 = lag11(score);
```

```
if person <= 300 then class=1;
if person >  300 then class=2;

* proc print data=d1;
*  var  time person class score score0 score1 score2 score3 score4
score5 score6
*  score7 score8 score9 score10 score11 ;
run;

data f1;
 set d1;
 if time = 11;
 file 'c:\mplus2\hba2pgms\data\c24s500.dat';
 put person 3.0 ' ' score0 8.4 score1 8.4 score2 8.4 score3 8.4 score4
8.4 score5 8.4 score6  8.4 score7  8.4 score8  8.4 score9  8.4 score10
8.4 score11 8.4 ;

* proc print data=f1;
*  var  time person class score score0 score1 score2 score3 score4
score5 score6
*  score7 score8 score9 score10 score11 ;

* proc means data=f1;
*  by class;
*  var score0 score1 score2 score3 score4 score5 score6 score7 score8
score9 score10 score11 ;
run;
```

DOS Batch Files Used in the Analysis of Simulated Data

One-class Model (Runk1.bat)

```
@echo off
rem*****************************************************************
rem*                           RUN CLASS 1 MODELS
rem*****************************************************************

set WORKING_DIR=C:\MPLUS2\HBA2PGMS

set DATA_DIR=C:\MPLUS2\HBA2PGMS\DATA

set FILES=*.dat

set RESULTS_DIR=C:\MPLUS2\HBA2PGMS

set EXE_FILE=C:\MPLUS2\MPDEVMC

set COMMON_DATA=C24SX.DAT

set INPUT_FILE=HBAXXXK1.PGM

set COMMON_DMP=HBAXXXK1.DAT

set ALL_DMP=HBAALLK1.DAT

set TMP_FILE=nc_all.tmp

set LOG_FILE=NOCONVK1.txt

set RUNONE=runone.bat
```

Two-class Model (Runk2.bat)

```
@echo off

rem*****************************************************************
rem *                        RUN Class 2 Models                   *
rem *****************************************************************

set WORKING_DIR=C:\MPLUS2\HBA2PGMS

set DATA_DIR=C:\MPLUS2\HBA2PGMS\DATA

set FILES=*.dat
```

```
set RESULTS_DIR=C:\MPLUS2\HBA2PGMS

set EXE_FILE=C:\MPLUS2\MPDEVMC

set COMMON_DATA=C24SX.DAT

set INPUT_FILE=HBAXXXK2.PGM

set COMMON_DMP=HBAXXXK2.DAT

set ALL_DMP=HBAALLK2.DAT

set TMP_FILE=nc_all.tmp

set LOG_FILE=NOCONVK2.txt

set RUNONE=runone.bat
```

Three-class Model (Runk3.bat)

```
@echo off

rem*****************************************************************
rem*                      RUN Class 3 Models                      *
rem ****************************************************************

set WORKING_DIR=C:\MPLUS2\HBA2PGMS

set DATA_DIR=C:\MPLUS2\HBA2PGMS\DATA

set FILES=*.dat

set RESULTS_DIR=C:\MPLUS2\HBA2PGMS

set EXE_FILE=C:\MPLUS2\MPDEVMC

set COMMON_DATA=C24SX.DAT

set INPUT_FILE=HBAXXXK3.PGM

set COMMON_DMP=HBAXXXK3.DAT

set ALL_DMP=HBAALLK3.DAT

set TMP_FILE=nc_all.tmp

set LOG_FILE=NOCONVK3.txt
```

```
set RUNONE=runone.bat
```

Runall.bat

```
@echo off

rem*******************************************************************
rem *                         RUNALL.BAT                            *
rem *******************************************************************
rem * Purpose: This is the top level file for the RUNALL utility. Type
rem *          RUNALL at the prompt.
rem * Written by:  Thuy Nguyen
rem *              Muthen & Muthen
rem *******************************************************************


rem*******************************************************************
rem *** DO NOT CHANGE ANYTHING BELOW THIS LINE IF YOU ARE NOT SURE WHY
rem *** CUSTOMIZED CONFIGURATIONS SHOULD BE SET IN RUNSTART.BAT
rem *******************************************************************

call runK1.bat
cd %DATA_DIR%
for %%a in (%FILES%) do call %WORKING_DIR%\%RUNONE% %%a
cd %WORKING_DIR%
call runend.bat

call runK2.bat
cd %DATA_DIR%
for %%a in (%FILES%) do call %WORKING_DIR%\%RUNONE% %%a
cd %WORKING_DIR%
call runend.bat

call runK3.bat
cd %DATA_DIR%
for %%a in (%FILES%) do call %WORKING_DIR%\%RUNONE% %%a
cd %WORKING_DIR%
call runend.bat
```

Runend.bat

```
@echo off

rem******************************************************************
rem *                            RUNEND.BAT
rem ******************************************************************
rem * Purpose:  The purpose of this file is to unset environment
rem *           variables that were set in RUNSTART.bat.  Nothing should
rem *           be changed in this file.  Check RUNSTART.bat
rem * Called:   This is called by RUNALL.bat
rem * Written by:  Thuy Nguyen
rem *              Muthen & Muthen
rem******************************************************************

set DATA_DIR=
set WORKING_DIR=
set RUNONE=
set FILES=
set EXE_FILE=
set COMMON_DATA=
set INPUT_FILE=
set COMMON_DMP=
set ALL_DMP=
set TMP_FILE=
set LOG_FILE=
```

Appendix D

SPSS Syntax Used to Examine Model Selection Criteria for Population Y

```
******************************************************************************.
*** READING IN Mplus OUTPUT FILES FOR CONDITION "HBA" (or Population Y) ***.
******************************************************************************.


*********************************.
***    TRUE ONE-CLASS MODEL ***.
*********************************.
DATA LIST FILE = 'C:\MPLUS3\hba3pgms\hbaoutKX.DAT'  FREE
 / REPLNUM KXCOND KXLL
   RVAR0 RVAR1 RVAR2 RVAR3 RVAR4 RVAR5 RVAR6 RVAR7 RVAR8 RVAR9 RVAR10 RVAR11
   KXB0 KXB1 KXVARB0 KXVARB1 .
FORMATS KXLL (F10.3).

LIST VAR = REPLNUM KXCOND KXLL KXB0 KXB1 .
FREQ VAR = KXLL /HISTOGRAM.

COMPUTE KXCLASS = 1.
COMPUTE FREEPARM = 16.
COMPUTE SAMPSIZE = 600 .
COMPUTE SAMPSIZ2 = (SAMPSIZE + 2) / 24.

COMPUTE LN_SAMP = LN(SAMPSIZE).
COMPUTE LN_SAMP2 = LN(SAMPSIZ2).

COMPUTE KXAIC = (-2*KXLL) + (2*FREEPARM).

COMPUTE KXBIC =  (-2*KXLL) + (FREEPARM*LN_SAMP).

COMPUTE KXABIC = (-2*KXLL) + (FREEPARM*LN_SAMP2).

LIST VAR =  REPLNUM KXCOND KXLL KXAIC KXBIC KXABIC.

SAVE OUTFILE = 'C:\MPLUS3\hba3pgms\hbaoutKX.SAV'
 /KEEP REPLNUM KXCOND KXLL KXAIC KXBIC KXABIC KXCLASS.
EXECUTE .


*********************************.
***    FALSE TWO-CLASS MODEL ***.
*********************************.
DATA LIST FILE = 'C:\MPLUS3\hba3pgms\hbaoutK0.DAT'  FREE
 / REPLNUM K0COND K0LL
   RVAR0 RVAR1 RVAR2 RVAR3 RVAR4 RVAR5 RVAR6 RVAR7 RVAR8 RVAR9 RVAR10 RVAR11
   KAB0 KAB1 K0VARB0 K0VARB1
   KBB0 KBB1 K0MEAN .
FORMATS K0LL (F10.3).

LIST VAR = REPLNUM K0COND K0LL KAB0 KAB1 KBB0 KBB1 K0MEAN.
FREQ VAR = K0LL /HISTOGRAM.

COMPUTE K0CLASS = 2.
COMPUTE FREEPARM = 19.
```

```
COMPUTE SAMPSIZE = 600 .
COMPUTE SAMPSIZ2 = (SAMPSIZE + 2) / 24.


COMPUTE LN_SAMP = LN(SAMPSIZE).
COMPUTE LN_SAMP2 = LN(SAMPSIZ2).


COMPUTE K0AIC = (-2*K0LL) + (2*FREEPARM).


COMPUTE K0BIC =  (-2*K0LL) + (FREEPARM*LN_SAMP).


COMPUTE K0ABIC = (-2*K0LL) + (FREEPARM*LN_SAMP2).


LIST VAR =  REPLNUM K0COND K0LL K0AIC K0BIC K0ABIC.


SAVE OUTFILE = 'C:\MPLUS3\hba3pgms\hbaoutK0.SAV'
 /KEEP REPLNUM K0COND K0LL K0AIC K0BIC K0ABIC K0CLASS.
EXECUTE .

******************************************************************.
***    ESTIMATING TYPE-I ERROR RATE FOR FALSE TWO-CLASS MODEL ***.
******************************************************************.
MATCH FILES
  FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTK0.SAV'
 /FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTKX.SAV'
 /BY REPLNUM .
EXECUTE .
FORMATS REPLNUM (F3.0) K0COND KXCOND (F1.0).

SELECT IF (K0COND GE 0 AND KXCOND EQ 0).
* LIST VAR = REPLNUM K0ABIC KXABIC .

COMPUTE ABICDIF = KXABIC - K0ABIC .
FREQ VAR = ABICDIF /HISTOGRAM.
* SORT CASES BY ABICDIF.
* LIST VAR = REPLNUM ABICDIF KXABIC K0ABIC .

COMPUTE BICDIF = KXBIC - K0BIC .
FREQ VAR = BICDIF /HISTOGRAM.

COMPUTE AICDIF = KXAIC - K0AIC .
FREQ VAR = AICDIF /HISTOGRAM.

COMPUTE POS_KXLL = -2*KXLL.
COMPUTE POS_K0LL = -2*K0LL.
COMPUTE LRTDIF = POS_KXLL - POS_K0LL .
FREQ VAR = LRTDIF /HISTOGRAM.
* SORT CASES BY LRTDIF.
* LIST VAR = REPLNUM LRTDIF POS_KXLL POS_K0LL .
```

```
***********************************************************************.
***     ESTIMATING TYPE-I ERROR RATE FOR FALSE THREE-CLASS MODEL ***.
***********************************************************************.
MATCH FILES
  FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTK2.SAV'
 /FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTK3.SAV'
 /BY REPLNUM .
EXECUTE .
FORMATS REPLNUM (F3.0) K2COND K3COND (F1.0).

SELECT IF (K2COND GE 0 AND K3COND EQ 0).
* LIST VAR = REPLNUM K2ABIC K3ABIC .

COMPUTE ABICDIF = K2ABIC - K3ABIC .
FREQ VAR = ABICDIF /HISTOGRAM.
SORT CASES BY ABICDIF.
LIST VAR = REPLNUM ABICDIF K2ABIC K3ABIC .

COMPUTE BICDIF = K2BIC - K3BIC .
FREQ VAR = BICDIF /HISTOGRAM.

COMPUTE AICDIF = K2AIC - K3AIC .
FREQ VAR = AICDIF /HISTOGRAM.

COMPUTE POS_K2LL = -2*K2LL.
COMPUTE POS_K3LL = -2*K3LL.
COMPUTE LRTDIF = POS_K2LL - POS_K3LL .
FREQ VAR = LRTDIF /HISTOGRAM.
SORT CASES BY LRTDIF.
LIST VAR = REPLNUM LRTDIF POS_K2LL POS_K3LL .


****************************************************.
***     FIRST, FOR THE FALSE ONE-CLASS MODELS  ***.
****************************************************.
DATA LIST FILE = 'C:\MPLUS3\hba3pgms\hbaoutK1.DAT'  FREE
 / REPLNUM K1COND K1LL
   RVAR0 RVAR1 RVAR2 RVAR3 RVAR4 RVAR5 RVAR6 RVAR7 RVAR8 RVAR9 RVAR10 RVAR11
   K1B0 K1B1 K1VARB0 K1VARB1 .
FORMATS K1LL (F10.3).
COMPUTE K1CLASS = 1.
COMPUTE FREEPARM = 16.
COMPUTE SAMPSIZE = 600 .
COMPUTE SAMPSIZ2 = (SAMPSIZE + 2) / 24.

COMPUTE LN_SAMP = LN(SAMPSIZE).
COMPUTE LN_SAMP2 = LN(SAMPSIZ2).

COMPUTE K1AIC = (-2*K1LL) + (2*FREEPARM).

COMPUTE K1BIC =  (-2*K1LL) + (FREEPARM*LN_SAMP).

COMPUTE K1ABIC = (-2*K1LL) + (FREEPARM*LN_SAMP2).
```

```
LIST VAR =  REPLNUM K1COND K1LL
   K1AIC K1BIC K1ABIC.

SAVE OUTFILE = 'C:\MPLUS3\hba3pgms\hbaoutK1.SAV'
 /KEEP REPLNUM K1COND K1LL K1AIC K1BIC K1ABIC K1CLASS.
EXECUTE .


***************************************************.
***     SECOND, FOR THE TRUE TWO-CLASS MODELS  ***.
***************************************************.
DATA LIST FILE = 'C:\MPLUS3\hba3pgms\hbaoutK2.DAT'  FREE
 / REPLNUM K2COND K2LL
   RVAR0 RVAR1 RVAR2 RVAR3 RVAR4 RVAR5 RVAR6 RVAR7 RVAR8 RVAR9 RVAR10 RVAR11
   K1B0 K1B1 K1VARB0 K1VARB1
   K2B0 K2B1 K1MEAN .
FORMATS K2LL (F10.3).
COMPUTE K2CLASS = 2.
COMPUTE FREEPARM = 19.
COMPUTE SAMPSIZE = 600 .
COMPUTE SAMPSIZ2 = (SAMPSIZE + 2) / 24.

COMPUTE LN_SAMP = LN(SAMPSIZE).
COMPUTE LN_SAMP2 = LN(SAMPSIZ2).

COMPUTE K2AIC = (-2*K2LL) + (2*FREEPARM).

COMPUTE K2BIC =  (-2*K2LL) + (FREEPARM*LN_SAMP).

COMPUTE K2ABIC = (-2*K2LL) + (FREEPARM*LN_SAMP2).

LIST VAR =  REPLNUM K2COND K2LL
  K2AIC K2BIC K2ABIC.

SAVE OUTFILE = 'C:\MPLUS3\hba3pgms\hbaoutK2.SAV'
 /KEEP REPLNUM K2COND K2LL K2AIC K2BIC K2ABIC K2CLASS.
EXECUTE .


***************************************************.
***     THIRD, FOR THE FALSE THREE-CLASS MODELS  ***.
***************************************************.
DATA LIST FILE = 'C:\MPLUS3\hba3pgms\hbaoutK3.DAT'  FREE
 / REPLNUM K3COND K3LL
   RVAR0 RVAR1 RVAR2 RVAR3 RVAR4 RVAR5 RVAR6 RVAR7 RVAR8 RVAR9 RVAR10 RVAR11
   K1B0 K1B1 K1VARB0 K1VARB1
   K2B0 K2B1 K3B0 K3B1  K1MEAN K2MEAN.
FORMATS K3LL (F10.3).
COMPUTE K3CLASS = 3.
COMPUTE FREEPARM = 22.
COMPUTE SAMPSIZE = 600 .
COMPUTE SAMPSIZ2 = (SAMPSIZE + 2) / 24.

COMPUTE LN_SAMP = LN(SAMPSIZE).
```

```
COMPUTE LN_SAMP2 = LN(SAMPSIZ2).

COMPUTE K3AIC = (-2*K3LL) + (2*FREEPARM).

COMPUTE K3BIC =  (-2*K3LL) + (FREEPARM*LN_SAMP).

COMPUTE K3ABIC = (-2*K3LL) + (FREEPARM*LN_SAMP2).

LIST VAR =  REPLNUM K3COND K3LL K3AIC K3BIC K3ABIC.
FREQ VAR = K3COND.

SAVE OUTFILE = 'C:\MPLUS3\hba3pgms\hbaoutK3.SAV'
 /KEEP REPLNUM K3COND K3LL K3AIC K3BIC K3ABIC K3CLASS.
EXECUTE .

***********************************************************.
****  PUTTING THEM ALL TOGETHER  ***********************.
***********************************************************.

MATCH FILES
  FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTK1.SAV'
 /FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTK2.SAV'
 /FILE = 'C:\MPLUS3\hba3PGMS\hbaOUTK3.SAV'
 /BY REPLNUM .
EXECUTE .
FORMATS REPLNUM (F3.0) K1COND K2COND K3COND (F1.0).


*************************************.
*** POWER and ACCURACY OF THE ABIC **.
*************************************.

IF (K1COND GT 0)K1ABIC=99999999.
IF (K2COND GT 0)K2ABIC=99999999.
IF (K3COND GT 0)K3ABIC=99999999.

COMPUTE MINABIC3 = MIN(K1ABIC, K2ABIC, K3ABIC).
COMPUTE ABIC_ACC = 0.
IF (MINABIC3 EQ K1ABIC)ABIC_ACC=1.
IF (MINABIC3 EQ K3ABIC)ABIC_ACC=3.
IF (MINABIC3 EQ K2ABIC)ABIC_ACC=2.
VAR LAB ABIC_ACC 'UNADJUSTED ACCURACY RATE'.
FREQ VAR = ABIC_ACC.

COMPUTE MINABIC2 = MIN(K1ABIC, K2ABIC) .
COMPUTE ABIC_PWR = 0.
IF (MINABIC2 EQ K1ABIC)ABIC_PWR=1.
IF (MINABIC2 EQ K3ABIC)ABIC_PWR=3.
IF (MINABIC2 EQ K2ABIC)ABIC_PWR=2.
VAR LAB ABIC_PWR 'UNADJUSTED POWER RATE'.
FREQ VAR = ABIC_PWR.

RECODE K1ABIC K2ABIC K3ABIC (99999999=SYSMIS).
COMPUTE ABIC_ACX=0.
```

```
IF (K1ABIC − K2ABIC GE −.85) AND (K2ABIC − K3ABIC LT −.85) ABIC_ACX=2.
VAR LAB ABIC_ACX 'TYPE−I ADJUSTED ACCURACY RATE'.
FREQ VAR = ABIC_ACX .

COMPUTE ABIC_PWX=0.
IF (K1ABIC − K2ABIC GE −.85) ABIC_PWX=2.
VAR LAB ABIC_PWX 'TYPE−I ADJUSTED POWER RATE'.
FREQ VAR = ABIC_PWX .

**************************************.
*** POWER and ACCURACY OF THE BIC **.
**************************************.

IF (K1COND GT 0)K1BIC=99999999.
IF (K2COND GT 0)K2BIC=99999999.
IF (K3COND GT 0)K3BIC=99999999.

COMPUTE MINBIC3 = MIN(K1BIC, K2BIC, K3BIC).
COMPUTE BIC_ACC = 0.
IF (MINBIC3 EQ K1BIC)BIC_ACC=1.
IF (MINBIC3 EQ K3BIC)BIC_ACC=3.
IF (MINBIC3 EQ K2BIC)BIC_ACC=2.
VAR LAB BIC_ACC 'UNADJUSTED ACCURACY RATE'.
FREQ VAR = BIC_ACC.

COMPUTE MINBIC2 = MIN(K1BIC, K2BIC) .
COMPUTE BIC_PWR = 0.
IF (MINBIC2 EQ K1BIC)BIC_PWR=1.
IF (MINBIC2 EQ K3BIC)BIC_PWR=3.
IF (MINBIC2 EQ K2BIC)BIC_PWR=2.
VAR LAB BIC_PWR 'UNADJUSTED POWER RATE'.
FREQ VAR = BIC_PWR.

RECODE K1BIC K2BIC K3BIC (99999999=SYSMIS).
COMPUTE BIC_ACX=0.
IF (K1BIC − K2BIC GE −10.37) AND (K2BIC − K3BIC LT −10.37) BIC_ACX=2.
VAR LAB BIC_ACX 'TYPE−I ADJUSTED ACCURACY RATE'.
FREQ VAR = BIC_ACX .

COMPUTE BIC_PWX=0.
IF (K1BIC − K2BIC GE −10.37) BIC_PWX=2.
VAR LAB BIC_PWX 'TYPE−I ADJUSTED POWER RATE'.
FREQ VAR = BIC_PWX .


**************************************.
*** POWER and ACCURACY OF THE AIC **.
**************************************.

IF (K1COND GT 0)K1AIC=99999999.
IF (K2COND GT 0)K2AIC=99999999.
IF (K3COND GT 0)K3AIC=99999999.

COMPUTE MINAIC3 = MIN(K1AIC, K2AIC, K3AIC).
```

109

```
COMPUTE AIC_ACC = 0.
IF (MINAIC3 EQ K1AIC)AIC_ACC=1.
IF (MINAIC3 EQ K3AIC)AIC_ACC=3.
IF (MINAIC3 EQ K2AIC)AIC_ACC=2.
VAR LAB AIC_ACC 'UNADJUSTED ACCURACY RATE'.
FREQ VAR = AIC_ACC.


COMPUTE MINAIC2 = MIN(K1AIC, K2AIC) .
COMPUTE AIC_PWR = 0.
IF (MINAIC2 EQ K1AIC)AIC_PWR=1.
IF (MINAIC2 EQ K3AIC)AIC_PWR=3.
IF (MINAIC2 EQ K2AIC)AIC_PWR=2.
VAR LAB AIC_PWR 'UNADJUSTED POWER RATE'.
FREQ VAR = AIC_PWR.


RECODE K1AIC K2AIC K3AIC (99999999=SYSMIS).
COMPUTE AIC_ACX=0.
IF (K1AIC - K2AIC GE 2.83) AND (K2AIC - K3AIC LT 2.83) AIC_ACX=2.
VAR LAB AIC_ACX 'TYPE-I ADJUSTED ACCURACY RATE'.
FREQ VAR = AIC_ACX .


COMPUTE AIC_PWX=0.
IF (K1AIC - K2AIC GE 2.83) AIC_PWX=2.
VAR LAB AIC_PWX 'TYPE-I ADJUSTED POWER RATE'.
FREQ VAR = AIC_PWX .



*************************************************.
*** POWER AND ACCURACY OF THE log-likelihood  ***.
*************************************************.

IF (K1COND GT 0)K1LL=99999999.
IF (K2COND GT 0)K2LL=99999999.
IF (K3COND GT 0)K3LL=99999999.


RECODE K1LL K2LL K3LL (99999999=SYSMIS).


DESC VAR = K1LL K2LL K3LL.


COMPUTE POS_K1LL = -2*K1LL.
COMPUTE POS_K2LL = -2*K2LL.
COMPUTE POS_K3LL = -2*K3LL.


COMPUTE DIF_12 = (POS_K1LL - POS_K2LL) .
COMPUTE DIF_23 = (POS_K2LL - POS_K3LL) .
FORMATS DIF_12 DIF_23  (F12.5).


COMPUTE SIG_12=0.
COMPUTE SIG_23=0.
IF (DIF_12 GT 7.81) SIG_12 = 1.
IF (DIF_23 LT 7.81) SIG_23 = 1.
VAL LAB SIG_12 SIG_23  0 'NO SIG DIF' 1 'SIG DIF' .


COMPUTE LL_PWR=0.
```

110

```
IF (SIG_12 = 1)LL_PWR=2.
VAR LAB LL_PWR 'POWER RATE USING CHI-SQUARE DIST' .
FREQ VAR = LL_PWR.

COMPUTE LL_ACC=0.
IF (SIG_12 = 1 AND SIG_23 = 1)LL_ACC=2.
VAR LAB LL_ACC 'ACCURACY RATE USING CHI-SQUARE DIST'.
FREQ VAR = LL_ACC.

COMPUTE SIG_12X=0.
COMPUTE SIG_23X=0.
IF (DIF_12 GT 8.82) SIG_12X = 1.
IF (DIF_23 LT 8.82) SIG_23X = 1.
VAL LAB SIG_12X SIG_23X  0 'NO SIG DIF' 1 'SIG DIF' .

COMPUTE LL_PWX=0.
IF (SIG_12X = 1)LL_PWX=2.
VAR LAB LL_PWX 'POWER RATE USING EMPIRICAL DIST' .
FREQ VAR = LL_PWX.

COMPUTE LL_ACX=0.
IF (SIG_12X = 1 AND SIG_23X = 1)LL_ACX=2.
VAR LAB LL_ACX 'ACCURACY RATE USING EMPIRICAL DIST'.
FREQ VAR = LL_ACX.

COMPUTE nchisqr=  -2*(K1LL - K2LL) - 3 .
desc var = nchisqr.

************************************************************************.
```

Mplus Input Programs Used in Monte Carlo Simulation of Evaluation Data
for Motivating Example

One-class GMM estimated from a one-class pseudo-population (K1$_T$)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO one class solution (tested against
         a true one class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=1;
            GCLASSES=1;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35);
            PATPROBS = 1.0;
            SAVE = c:\mplus3\MACRO\REP_K1T.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K1T.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0];
            intercpt*70.751 slope*2.463;
            intercpt with slope@-4.414;
            del1*34.629 del2*33.937 del3*34.274 del4*30.307 del5*34.722;

             %c#1%
             [intercpt*67.851 slope*-1.286];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

             %c#1%
             [intercpt*67.851 slope*-1.286];


OUTPUT:     TECH1 TECH9;
```

One-class GMM estimated from a two-class pseudo-population (K1$_F$)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO two class solution (tested against
         a false one class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=1;
            GCLASSES=2;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35);
            PATPROBS = 1.0;
            SAVE = c:\mplus3\MACRO\REP_K1F.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K1F.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0];
            intercpt*33.762 slope*2.547;
            intercpt with slope@-5.518;
            del1*32.968 del2*34.587 del3*35.055 del4*31.227 del5*32.447;

             %c#1%
            [intercpt*61.456 slope*-1.436];

             %c#2%
            [intercpt*73.737 slope*-1.174];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

             %c#1%
            [intercpt*67.851 slope*-1.286];

OUTPUT:     TECH9;
```

Two-class GMM estimated from a two-class pseudo-population (K2_T)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO two class solution (tested against
         a true two class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=2;
            GCLASSES=2;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35);
            PATPROBS = 1.0 ;
            SAVE = c:\mplus3\MACRO\REP_K2T.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K2T.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0];
            intercpt*33.762 slope*2.547;
            intercpt with slope@-5.518;
            del1*32.968 del2*34.587 del3*35.055 del4*31.227 del5*32.447;

          [c#1@-0.090];

            %c#1%
          [intercpt*61.456 slope*-1.436];

            %c#2%
          [intercpt*73.737 slope*-1.174];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

            %c#1%
          [intercpt*63.645 slope*-2.071];

            %c#2%
          [intercpt*72.057 slope*-0.501];

OUTPUT:     TECH9;
```

114

Two-class GMM estimated from a one-class pseudo-population (K2$_F$)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO one class solution (tested against
         a false two class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=2;
            GCLASSES=1;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35);
            PATPROBS = 1.0;
            SAVE = c:\mplus3\MACRO\REP_K2F.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K2F.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0];
            intercpt*70.751 slope*2.463;
            intercpt with slope@-4.414;
            del1*34.629 del2*33.937 del3*34.274 del4*30.307 del5*34.722;

             %c#1%
            [intercpt*67.851 slope*-1.286];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

             %c#1%
            [intercpt*63.645 slope*-2.071];

             %c#2%
            [intercpt*72.057 slope*-0.501];

OUTPUT:     TECH1 TECH9;
```

Two-class GMM estimated from a three-class pseudo-population (K2FF)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO THREE class solution (tested against
         a TRUE three class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=2;
            GCLASSES=3;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35) ;
            PATPROBS = 1.0 ;
            SAVE = c:\mplus3\MACRO\REP_K3FF.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K3FF.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0];
            intercpt*39.744 slope*1.433;
            intercpt with slope@-7.226;
            del1*34.144 del2*34.200 del3*35.884 del4*31.476 del5*29.946;

           [c#1@2.138];
           [c#2@2.259];

             %c#1%
           [intercpt*61.614 slope*-1.817];

             %c#2%
           [intercpt*72.875 slope*-1.260];

             %c#3%
           [intercpt*72.794 slope*2.973];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

             %c#1%
           [intercpt*63.645 slope*-2.071];

             %c#2%
           [intercpt*72.057 slope*-0.501];

OUTPUT:     TECH9;
```

116

Three-class GMM estimated from a two-class pseudo-population (K3$_F$)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO two class solution (tested against
         a false three class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=3;
            GCLASSES=2;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35);
            PATPROBS = 1.0 ;
            SAVE = c:\mplus3\MACRO\REP_K3F.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K3F.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
             intercpt BY del1-del5@1;
             slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
             [del1-del5@0];
             intercpt*33.762 slope*2.547;
             intercpt with slope@-5.518;
             del1*32.968 del2*34.587 del3*35.055 del4*31.227 del5*32.447;

              [c#1@-0.090];

              %c#1%
             [intercpt*61.456 slope*-1.436];

              %c#2%
             [intercpt*73.737 slope*-1.174];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

              %c#1%
             [intercpt*63.645 slope*-2.071];

              %c#2%
             [intercpt*67.851 slope*-1.286];

              %c#3%
             [intercpt*72.057 slope*-0.501];

OUTPUT:     TECH9;
```

Three-class GMM estimated from a three-class pseudo-population (K3$_T$)

```
TITLE:   Monte Carlo simulation of GMM that simulates conditions
         specified by MACRO THREE class solution (tested against
         a TRUE three class model)

MONTECARLO: NAMES ARE del1-del5;
            NOBSERVATIONS = 442;
            NREPS = 500;
            SEED = 376568507;
            NCLASSES=3;
            GCLASSES=3;
            PATMISS = del1(0.0) del2 (.10) del3 (.11) del4 (.19) del5 (.35) ;
            PATPROBS = 1.0 ;
            SAVE = c:\mplus3\MACRO\REP_K3T.DAT;
            RESULTS = c:\mplus3\MACRO\OUT_K3T.DAT;

ANALYSIS:   TYPE = MIXTURE MISSING;
            MITERATIONS = 5000;

MODEL MONTECARLO:

            %OVERALL%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0];
            intercpt*39.744 slope*1.433;
            intercpt with slope@-7.226;
            del1*34.144 del2*34.200 del3*35.884 del4*31.476 del5*29.946;

          [c#1@2.138];
          [c#2@2.259];

            %c#1%
          [intercpt*61.614 slope*-1.817];
            %c#2%
          [intercpt*72.875 slope*-1.260];
            %c#3%
          [intercpt*72.794 slope*2.973];

MODEL: %overall%
            intercpt BY del1-del5@1;
            slope BY del1@0 del2@1 del3@2 del4@3 del5@4 ;
            [del1-del5@0]; intercpt with slope*;

            %c#1%
          [intercpt*63.645 slope*-2.071];
            %c#2%
          [intercpt*67.851 slope*-1.286];
            %c#3%
          [intercpt*72.057 slope*-0.501];

OUTPUT:     TECH9;
```

Tables

Table 1. *Monte Carlo simulation studies of latent class enumeration (in chronological order)*

Study	Type of Model	Hypotheses Tested	Model selection criteria
Wolfe (1971)	Latent class model	True H_0: $k_0 = 1$ False H_A: $k_1 = k_0 > 1$	LRT
Everitt (1981, 1988)	Latent class model	True H_0: $k_0 = 1$ False H_A: $k_1 = 2$	LRT
Mendell, Thode, and Finch (1991)	Latent class model	False H_0: $k_0 = 1$ True H_A: $k_1 = 2$	LRT
Mendell, Finch, and Thode (1993)	Latent class model	False H_0: $k_0 = 1$ True H_A: $k_1 = 2$	Engelman-Hartigan test, Grubb test, Dixon's test Anderson-Darling test, Craver-Von Mises test, Wilk-Shapiro test, skewness test, kurtotsis test, range test, Geary test, Filliben's probability plot correction test, D'Agostino test, LRT.
Gutierrez, Wang, Lee, and Taylor (1995)	Latent class model	True H_0: $k_0 = 1$ False H_A: $k_1 = 2$	LRT
Lin and Dayton (1997)	Latent class model	False H_0: $k_0 = 1$ True H_A: $k_1 = 2$	BIC, AIC, CAIC, LRT,
Yang (1999)	Latent class model	True H_0: $k_0 = 1$ False H_A: $k_1 = 2, 3, 4$	BIC, AIC, LRT, ABIC, DBIC, HQIC, HT-AIC
		False H_0: $k_0 = 1, 3, 4$ True H_A: $k_1 = 2$	
		False H_0: $k_0 = 1, 2, 4$ True H_A: $k_1 = 3$	
		False H_0: $k_0 = 1, 2, 3$ True H_A: $k_1 = 4$	
Masyn and Brown (2001)	Growth mixture model	True H_0: $k_0 = 1$ False H_A: $k_1 = 2$	BIC, AIC, ABIC, LRT, CLC, NEC, ICL-BIC
		False H_0: $k_0 = 1, 3$ True H_A: $k_1 = 2$	

Continued on the next page (see *Note* on next page).

Table 1. *Monte Carlo simulation studies of latent class enumeration (continued)*

Study	Type of Model	Hypotheses Tested	Examined model selection criteria
Masyn (2001)	Growth mixture model	True H_0: $k_0 = 1$ False H_A: $k_1 = 2, 3$	BIC, AIC, ABIC, LRT, CLC, NEC, ICL-BIC
		False H_0: $k_0 = 1, 3$ True H_A: $k_1 = 2$	
		False H_0: $k_0 = 1, 2$ True H_A: $k_1 = 3$	
Lo, Mendell, and Rubin (2001)	Latent class model	False H_0: $k_0 = 1$ True H_A: $k_1 = 2$	LRT
		False H_0: $k_0 = 2$ True H_A: $k_1 = 3$	

Note. BIC = Bayesian Information Criterion, AIC = Akaike Information Criterion, ABIC = Sample-size adjusted BIC, LRT = Likelihood Ratio Test, CAIC = Bozdogan's Consistnet AIC, DBIC = Draper's modified BIC, HQIC = Hannan and Quinn's Information Criterion, HT-AIC = Hurvich and Tsai's AIC, CLC = Classification Likelihood Criterion, NEC = Normalized Entropy Criterion, ICL-BIC = Integrated Classification Likelihood BIC.

Table 2. *Applications of growth mixture modeling (ordered alphabetically)*

Study	Outcome	Sample	Finding
Colder et al. (2001)	Cigarette smoking	N = 323 adolescents	BIC identified five latent trajectory classes.
Colder, Campbell, Ruel, Richardson, & Flay (2002)	Alcohol use	N = 1,918 adolescents	BIC identified four-class model for alcohol quantity, three-class model for alcohol frequency, and 5-class model for alcohol quantity-frequency measure.
Dedrick, Greenbaum, & Vaughn (2001)	Reading achievement	N = 100 second-grade students	BIC identified two latent trajectory classes.
Greenbaum, Dedrick, & Brown (2002)	Delinquent behavior	N = 442 children	BIC and ABIC identified four-class model; AIC identified five-class model.
Greenbaum, Del Boca, Darkes, & Goldman (2001)	Alcohol use	N = 229 college freshmen	BIC indicated four-class model; AIC and ABIC indicated five-class model.
Li, Duncan, & Hops (2001)	Alcohol use	N = 179 middle and high school students	BIC indicated one-class model; AIC, ABIC, and LRT indicated two-class model.
Muthén et al. (2002)	Aggressive behavior	N = 199 first- and second grade boys	BIC indicated preference for 4-class model.
Muthén, Khoo, Francis, & Boscardin (2000)	Reading skills	N = 409 Kindergarten-age children	LRT identified four latent trajectory classes.
Oxford, Gilchrist, Morrison, Gillmore, Lohr, & Lewis (2003)	Alcohol use	N = 215 adolescent females	BIC identified two-class model for alcohol quantity, four-class model for alcohol frequency.
Petras et al. (2001)	School removal	N = 200 boys	BIC indicated three-class model; AIC and ABIC indicated four-class model.
Tucker, Orlando, & Ellickson (2003)	Binge drinking	N = 5,694 adolescents	BIC identified six-class model.

Note. BIC = Bayesian Information Criterion, AIC = Akaike Information Criterion, ABIC = Sample-size adjusted BIC, LRT = Likelihood Ratio Test.

Table 3. *Growth parameter values, variances, and covariances for simulated data*

Population/Design	Class 1		Class 2		I with S covariance[a]	Level 1 variance
	Intercept (Variance)	Slope (Variance)	Intercept (Variance)	Slope (Variance)		
A	100.00 (57.0)	0.00 (3.00)	97.74 (57.00)	-0.52 (3.00)	0.00	40.0
B, L, O, R	100.00 (57.0)	0.00 (3.00)	96.23 (57.00)	-0.87 (3.00)	0.00	40.0
C, J, M, P, S	100.00 (57.0)	0.00 (3.00)	93.96 (57.00)	-1.38 (3.00)	0.00	40.0
D, K, N, Q, T, U, V, W, X, Y, Z, AA, AB, AC, AD, AE, AF	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	0.00	40.0
E	100.00 (57.0)	0.00 (3.00)	90.94 (57.00)	-2.08 (3.00)	0.00	40.0
F	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	0.00	90.0
G	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	0.00	60.0
H	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	0.00	26.0
I	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	0.00	15.0
AG	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	3.92	40.0
AH	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	7.84	40.0
AI	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	-3.92	40.0
AJ	100.00 (57.0)	0.00 (3.00)	92.45 (57.00)	-1.73 (3.00)	-7.84	40.0

Note. Letters in Population/Design column represent the 36 unique combinations of population and design characteristics (see Table 4 for experimental conditions). I = intercept. S = slope. [a]Covariances between intercepts and slopes were class-invariant for all experimental conditions.

Table 4. *Population/designs and experimental conditions*

Population / Design	Experimental condition					
	Standardized between-class difference in growth parameters	Percentage of variance explained by growth parameters	Correlation between intercept and slope	Sample size	Number of repeated measures	Missing data pattern
A	.30[a]	60	@.00	600	6	none
B	.50[a]	60	@.00	600[c]	6	none
C	.80[a]	60	@.00	600[b]	6	none
D	1.00[a]	60[a]	@.00	600[a]	6[a]	none[a]
E	1.20[a]	60	@.00	600	6	none
F	1.00	40[a]	@.00	600	6	none
G	1.00	50[a]	@.00	600	6	none
H	1.00	70[a]	@.00	600	6	none
I	1.00	80[a]	@.00	600	6	none
J	.80	60	@.00	200[b]	6	none
K	1.00	60	@.00	200[a]	6	none
L	.50	60	@.00	400[c]	6	none
M	.80	60	@.00	400[b]	6	none
N	1.00	60	@.00	400[a]	6	none
O	.50	60	@.00	800[c]	6	none
P	.80	60	@.00	800[b]	6	none
Q	1.00	60	@.00	800[a]	6	none
R	.50	60	@.00	1000[c]	6	none
S	.80	60	@.00	1000[b]	6	none
T	1.00	60	@.00	1000[a]	6	none
U	1.00	60	@.00	600	3[a]	none

Continued on the next page (see *Note* on next page).

Table 4. *Population/designs and experimental conditions (continued)*

Population / Design	Experimental condition					
	Standardized between-class difference in growth parameters	Percentage of variance explained by growth parameters	Correlation between intercept and slope	Sample size	Number of repeated measures	Missing data pattern
V	1.00	60	@.00	600	4[a]	none[b]
W	1.00	60	@.00	600	5[a]	none
X	1.00	60	@.00	600	9[a]	none
Y	1.00	60	@.00	600	12[a]	none[a,b]
Z	1.00	60	@.00	600	12	every other time point missing 25%[a]
AA	1.00	60	@.00	600	12	every other time point missing 50%[a]
AB	1.00	60	@.00	600	12	every other time point missing 75%[a]
AC	1.00	60	@.00	600	12	every 2nd and 3rd time point missing 25%[b]
AD	1.00	60	@.00	600	12	every 2nd and 3rd time point missing 50%[b]
AE	1.00	60	@.00	600	12	every 2nd and 3rd time point missing 75%[b]
AF	1.00	60	*.00[a]	600	6	none
AG	1.00	60	*.30[a]	600	6	none
AH	1.00	60	*.60[a]	600	6	none
AI	1.00	60	*-.30[a]	600	6	none
AJ	1.00	60	*-.60[a]	600	6	none

Note. Letters in Population/Design column represent the 36 unique combinations of population and design characteristics (see Table 3 for population characteristics). Cells within columns with same superscripted letters indicate population/designs used for comparison of that experimental condition holding other conditions constant. @ = correlation fixed at value in data generation model and not estimated in estimation model. * = correlation fixed at value in data generation model and freely estimated in estimation model.

Table 5. *Planned missingness designs*

Missingness Pattern						Time						
	1	2	3	4	5	6	7	8	9	10	11	12
No missing data	X	X	X	X	X	X	X	X	X	X	X	X
Design 1	X		X		X		X		X		X	
Design 2	X			X			X			X		

Note. Design 1 indicates planned missingness at every other time point. Design 2 indicates planned missingness for every 2nd and 3rd time point. X = No missing data for that time point.

Table 6. *95% Confidence intervals for p* = .50 and .90*

Number of Replications	p^*	
	.50	.90
100	$\pm$ .100	$\pm$ .060
200	$\pm$ .071	$\pm$ .042
300	$\pm$ .058	$\pm$ .035
400	$\pm$ .050	$\pm$ .030
500	$\pm$ .045	$\pm$ .027

Note. p^* = estimated level of power.

Table 7. *Convergence rates by model and experimental condition*

Population / Design	Model				
	K1$_T$	K1$_F$	K2$_T$	K2$_F$	K3$_F$
A	100	100	100	99.4	99.0
B	100	100	99.8	99.2	98.2
C	100	100	99.6	99.8	98.8
D	100	100	99.4	99.8	98.2
E	100	100	99.6	99.8	98.4
F	100	100	99.6	99.2	97.4
G	100	100	100	99.8	98.0
H	100	100	100	99.4	98.8
I	100	100	100	100	98.8
J	100	100	100	100	99.4
K	100	100	100	99.8	98.4
L	100	100	100	99.6	98.4
M	100	100	99.8	99.4	98.8
N	100	100	99.8	99.8	99.2
O	100	100	99.9	99.6	98.9
P	100	100	100	99.8	98.6
Q	100	100	100	99.4	99.0
R	100	100	99.2	99.8	99.2
S	100	100	100	99.8	98.2
T	100	100	100	99.6	98.6
U	100	100	100	99.8	98.6
V	100	100	99.8	99.4	98.2
W	100	100	100	100	99.0
X	100	100	100	99.2	99.2
Y	100	100	100	100	97.2
Z	100	100	99.8	97.4	91.2

Table 7. *Convergence rates by model and experimental condition (continued)*

Population / Design	Model				
	K1$_T$	K1$_F$	K2$_T$	K2$_F$	K3$_F$
AA	100	100	98.4	98.0	94.4
AB	100	100	100	96.8	92.8
AC	100	100	100	98.0	94.0
AD	100	100	100	98.4	94.4
AE	100	100	100	97.4	93.8
AF	100	100	99.6	99.6	97.6
AG	100	100	97.6	97.4	89.2
AH	100	100	99.8	99.0	98.4
AI	100	100	100	99.4	99.0
AJ	100	100	100	98.2	98.6
Mean	100	100	99.8	99.2	97.4
Range	100 to 100	100 to 100	97.6 to 100	96.8 to 99.6	89.2 to 99.0

Note. Values indicate percentage of 500 replications that converged within 5000 iterations. K1$_T$ = One-class model estimated from a one-class pseudo-population. K1$_F$ = One-class model estimated from a two-class pseudo-population. K2$_T$ = Two-class model estimated from a two-class pseudo-population. K2$_F$ = Two-class model estimated from a one-class pseudo-population. K3$_F$ = Three-class model estimated from a two-class pseudo-population. See Tables 3 and 4 for a description of Population/Design codes.

Table 8. *Effective Type I error rates for K1$_T$ versus K2$_F$ models (averaged within experimental condition and quartiles of λ^2)*

Experimental condition / λ^2	Model selection criterion			
	AIC	BIC	ABIC	LRT
Sample size (*d* = .50)	.160	<.001	.032	.076
Sample size (*d* = .80)	.168	.001	.056	.070
Sample size (*d* = 1.0)	.147	.001	.049	.072
Number of repeated measures	.168	<.001	.038	.081
Between-class differences in growth parameters	.162	<.001	.032	.071
Percentage of variance explained	.172	.002	.037	.086
Correlation between growth parameters	.212	<.001	.052	.100
$\lambda^2 < 2.60$	.177	.001	.054	.083
$2.61 < \lambda^2 < 8.00$	.167	.001	.058	.079
$8.01 < \lambda^2 < 17.00$	.168	<.001	.032	.081
$17.01 < \lambda^2$	.152	<.001	.035	.086
Overall	.167	.001	.046	.082

Note. λ^2 = noncentrality parameter based on chi-square distribution with df = 3. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion. ABIC = Sample size adjusted BIC. LRT = Likelihood Ratio Test. *d* = standardized between-class difference in growth parameters.

Table 9. *Effective Type I error rates for K2$_T$ versus K3$_F$ models (averaged within experimental condition and quartiles of λ^2)*

Experimental condition / λ^2	Model selection criterion			
	AIC	BIC	ABIC	LRT
Sample size (*d* = .50)	.199	.001	.047	.104
Sample size (*d* = .80)	.217	.000	.078	.102
Sample size (*d* = 1.0)	.220	.000	.077	.106
Number of repeated measures	.208	.001	.049	.099
Between-class differences in growth parameters	.203	.000	.050	.100
Percentage of variance explained	.234	.000	.065	.113
Correlation between growth parameters	.244	.000	.064	.141
$\lambda^2 < 2.60$	.200	.001	.068	.101
$2.61 < \lambda^2 < 8.00$	.214	.000	.080	.104
$8.01 < \lambda^2 < 17.00$	.226	.000	.052	.107
$17.01 < \lambda^2$	.210	.000	.045	.110
Overall	.213	.001	.062	.105

Note. λ^2 = noncentrality parameter based on chi-square distribution with df = 3. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion. ABIC = Sample size adjusted BIC. LRT = Likelihood Ratio Test. *d* = standardized between-class difference in growth parameters.

Table 10. *95% critical values for derivation of Type I error rates obtained from empirical distributions*

Population / Design	Model selection criterion								
	K1$_T$ versus K2$_F$					K2$_T$ versus K3$_F$			
	AIC	BIC	ABIC	LRT		AIC	BIC	ABIC	LRT
A	3.15	-10.04	-0.52	9.15		2.61	-10.58	-1.06	8.61
B	2.27	-10.92	-1.39	8.27		3.64	-9.55	-0.03	9.64
C	2.22	-10.95	-1.45	8.22		3.36	-9.98	-0.30	9.36
D	2.31	-10.88	-1.36	8.31		4.18	-9.01	0.51	10.18
E	3.01	-10.18	-0.66	9.01		3.61	-9.58	-0.06	9.61
F	2.62	-10.57	-1.04	8.62		4.48	-8.72	0.81	10.48
G	3.96	-9.23	0.30	9.96		4.13	-9.06	0.05	10.13
H	3.34	-9.85	-0.33	9.34		3.75	-9.41	0.11	9.77
I	2.72	-10.47	-0.95	8.72		4.40	-8.79	0.80	10.41
J	3.94	-5.95	3.54	9.94		4.43	-5.47	4.04	10.43
K	2.71	-7.18	2.32	8.71		3.50	-6.40	3.10	9.50
L	2.86	-9.11	0.41	8.86		3.63	-8.34	1.18	9.63
M	3.27	-8.71	0.82	9.27		3.20	-8.78	0.74	9.20
N	2.56	-9.42	0.10	8.56		3.69	-8.28	1.24	9.69
O	3.21	-10.84	-1.31	9.21		2.85	-11.20	-1.69	8.85
P	3.55	-10.50	-0.98	9.55		3.60	-10.46	-0.93	9.60
Q	1.99	-12.06	-2.54	7.99		2.91	-11.15	-1.62	8.91
R	3.07	-11.66	-2.13	9.07		3.56	-11.17	-1.64	9.56
S	2.45	-12.28	-2.75	8.45		3.71	-11.01	-1.48	9.71
T	3.11	-11.62	-2.09	9.11		3.82	-10.91	-1.38	9.82
U	3.64	-9.55	0.00	9.64		2.94	-10.25	-0.73	8.94
V				same as "Y"					
W				same as "Y"					
X	2.98	-10.21	-0.68	8.98		3.23	-9.96	-0.43	9.23
Y	2.83	-10.37	-0.85	8.82		2.83	-10.37	-0.85	8.82
Z				same as "Y"					
AA				same as "Y"					
AB				same as "Y"					
AC				same as "Y"					
AD				same as "Y"					
AE				same as "Y"					
AF	4.05	-9.16	0.36	10.05		4.03	-9.16	0.38	10.03
AG				same as "AF"					
AH				same as "AF"					
AI				same as "AF"					
AJ				same as "AF"					

Note. AIC = Akaike Information Criterion. BIC = Bayesian Information Criterion. ABIC = Sample size adjusted BIC. LRT = Likelihood Ratio Test. See Tables 3 and 4 for a description of Population/Design codes.

Table 11. *Estimates of statistical power, accuracy, and population/design by λ^2*

λ^2	Power estimated from noncentral chisquare distribution[a]	Power estimated from Monte Carlo simulation	Accuracy estimated from Monte Carlo simulation	Population / Design
0.745	.098	.052	.050	A
0.918	.110	.060	.052	L
1.304	.138	.046	.040	AH
1.454	.150	.142	.138	B
1.714	.170	.124	.118	O
1.889	.184	.112	.108	R
1.898	.184	.084	.074	U
2.460	.230	.162	.150	AF
2.582	.240	.120	.112	J
3.371	.306	.142	.098	AG
3.672	.331	.180	.168	AI
3.875	.348	.234	.216	M
4.388	.391	.318	.298	K
4.412	.393	.334	.298	V
4.966	.438	.408	.386	C
5.948	.515	.476	.444	F
6.745	.572	.450	.416	P
7.832	.644	.614	.588	N
8.033	.656	.612	.566	W
8.149	.663	.624	.574	S
8.322	.674	.564	.530	G
11.412	.820	.824	.772	D
12.270	.849	.732	.686	AJ
14.290	.903	.854	.806	H
15.297	.923	.928	.872	Q
15.996	.934	.861	.802	AE
16.547	.942	.906	.798	AD
17.875	.958	.966	.908	X
17.888	.958	.934	.804	AB
18.545	.964	.952	.890	T
18.650	.965	.958	.846	AC
18.670	.965	.950	.890	I
18.799	.966	.949	.836	AA
19.270	.970	.962	.836	Z
20.745	.979	.954	.878	Y
22.377	.987	.986	.922	E

Note. Estimates based on distributions with alpha = .05. See Tables 3 and 4 for a description of Population/Design codes. [a]df = 3.

Table 12. *Model selection information for evaluation data and corresponding Monte Carlo simulations[a]*

Model	Model selection information				
	AIC	BIC	ABIC	LL	Number of freely estimated parameters
Evaluation data					
One-class	13093.95	13134.86	13103.13	-6536.98	10
Two-class	13084.51	13137.69	13096.44	-6529.25	13
Three-class	13049.50	13114.96	13064.18	-6508.75	16
Simulated data[a]					
One-class	13098.51	13139.42	13107.68	-6539.25	10
Two-class	13083.34	13136.53	13095.27	-6528.67	13
Three-class	13049.00	13114.46	13063.69	-6510.21	16

Note. [a]Mean values based on 500 replications. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; ABIC = Sample size adjusted BIC; LL = Log Likelihood.

Table 13. *Parameter estimates for one-, two-, and three-class growth mixture models of evaluation data*

Parameter	Model		
	One-class	Two-class	Three-class
Intercept			
class 1	67.851	61.456	61.614
class 2	na	73.737	72.875
class 3	na	na	72.794
Linear slope			
class 1	-1.286	-1.436	-1.817
class 2	na	-1.174	-1.260
class 3	na	na	2.973
Variance			
Intercept	70.751	33.762	39.744
Slope	2.463	2.547	1.433
Covariance			
intercept with slope	-4.414	-5.518	-7.226
Residual Variance			
time 1	34.629	32.968	34.144
time 2	33.937	34.587	34.200
time 3	34.274	35.055	35.884
time 4	30.307	31.227	31.476
time 5	34.722	32.447	29.946
Latent trajectory class mean			
class 1	0.000	-0.090	2.138
class 2	na	0.000	2.259
class 3	na	na	0.000

Note. na = not applicable.

Figures
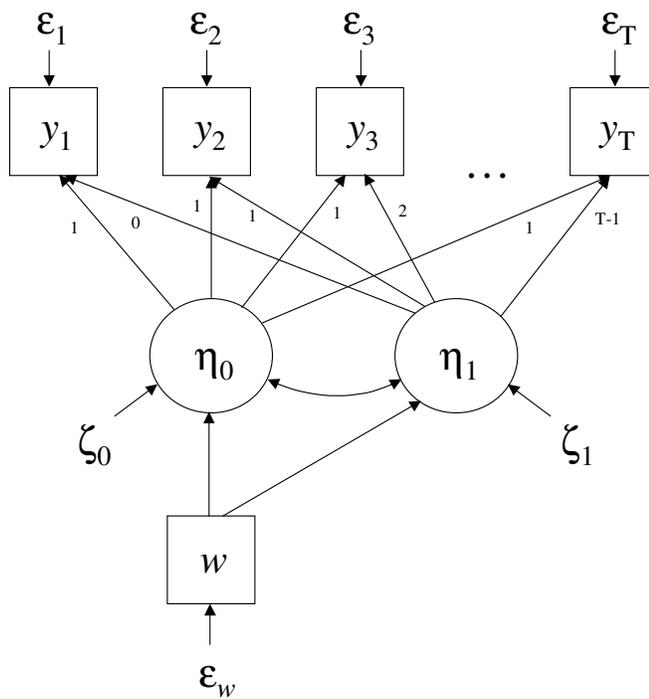
Figure 1. Latent growth model with indicator variables (y_t), growth factors (η_0 and η_1), and time-invariant covariate (w).
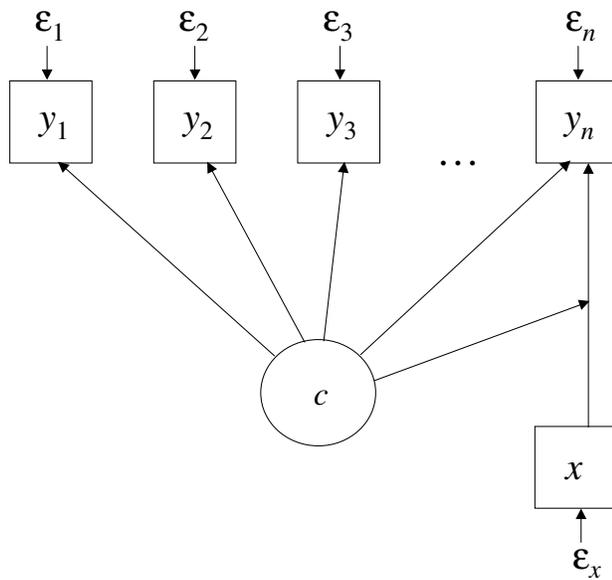
Figure 2. Latent variable mixture model with indicator variables (y_n), latent class variable (c), and covariate (x).
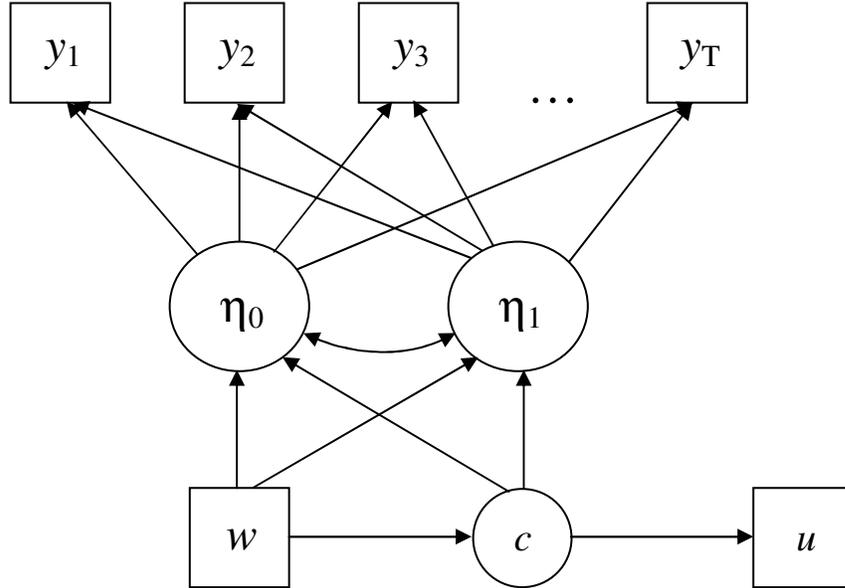
Figure 3. General growth mixture model with indicator variables (y_t), growth factors (η_0 and η_1), time-invariant covariate (w), latent class variable (c), and outcome variable (u) (factor loadings and error terms omitted).
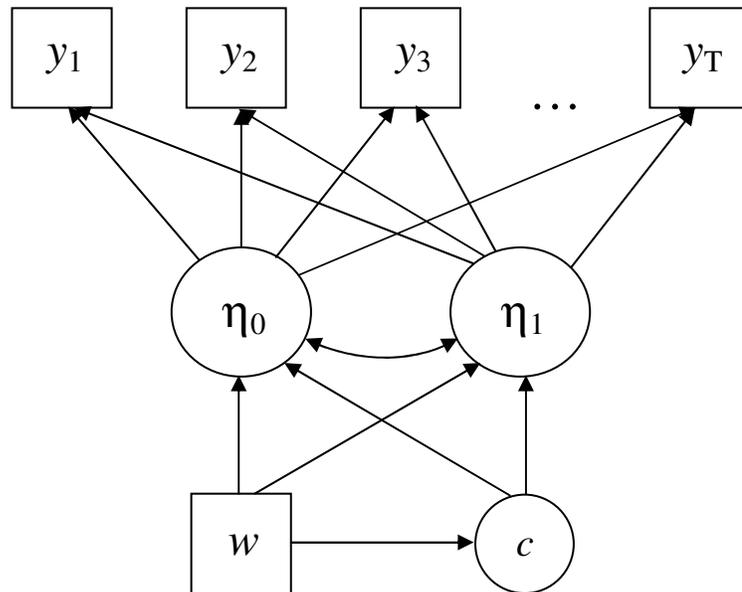
Figure 4. Growth mixture model with indicator variables (y_t), growth factors (η_0 and η_1), time-invariant covariate (w), and latent class variable (c) (factor loadings and error terms omitted).

138

Class 1

Class 2
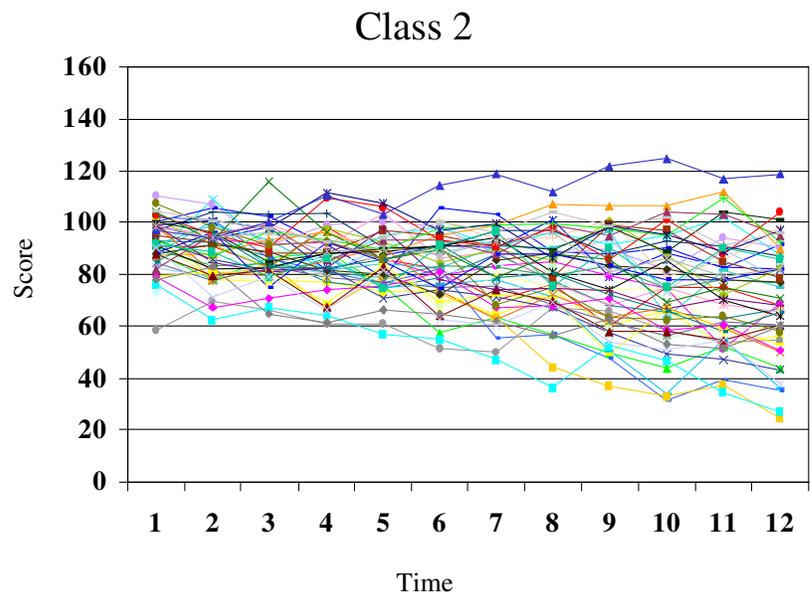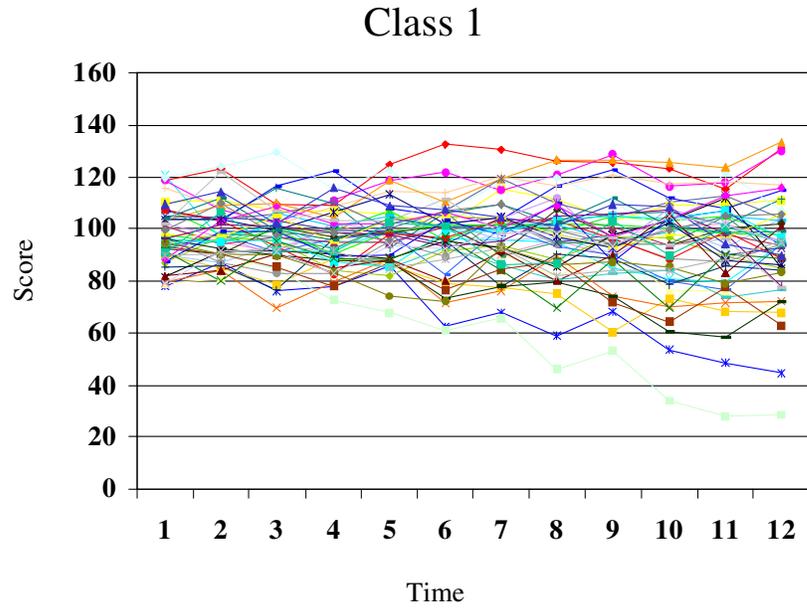
Figure 5. Spaghetti plots for simulated data from Population Y (*n* = 50 each).

139
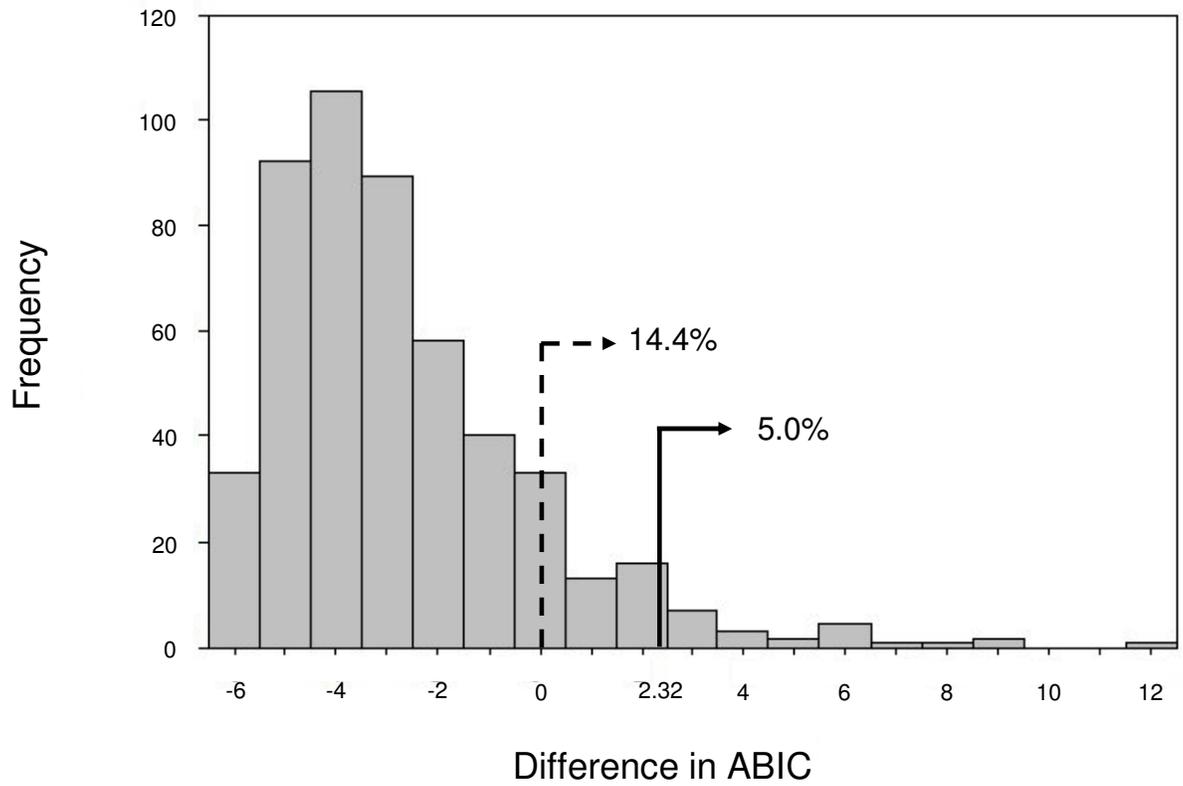
Figure 6. Frequency distribution for $K1_T - K2_F$ ABIC values for Population K (percentages indicate percentage of replications greater than corresponding difference between $K1_T - K2_F$ models).
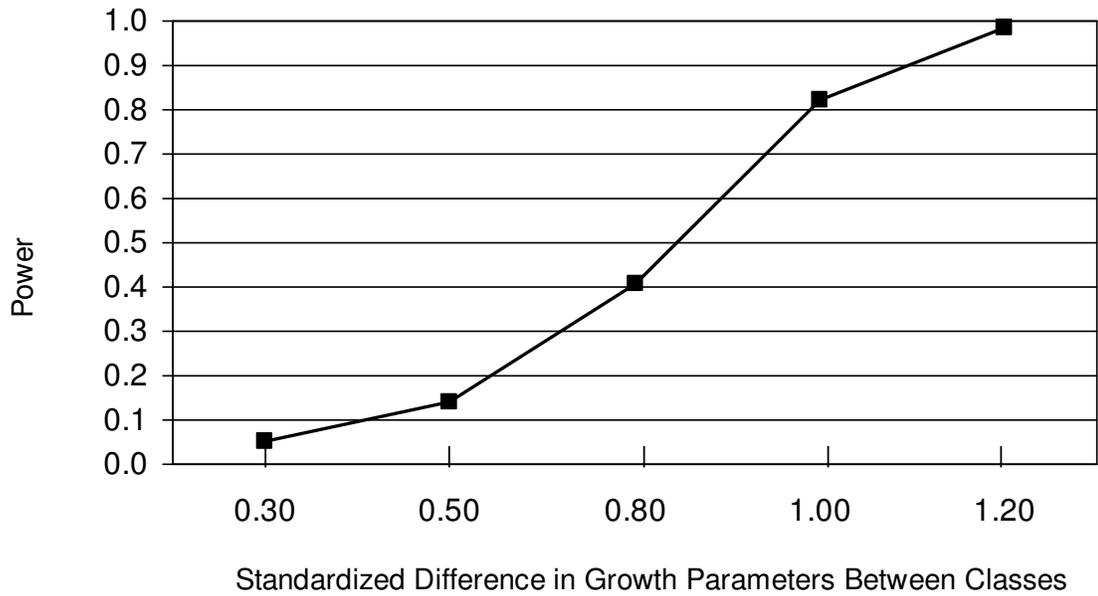
Figure 7. Adjusted power estimates for standardized between-class differences in growth parameters.
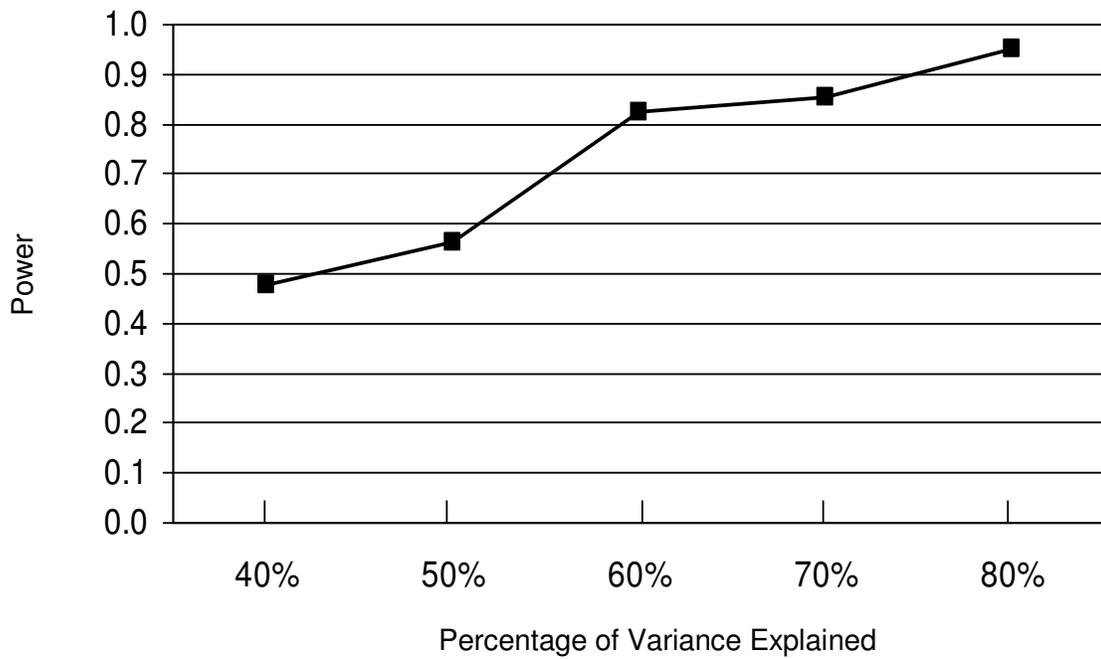
Figure 8. Adjusted power estimates for percentage of variance explained by growth parameters.
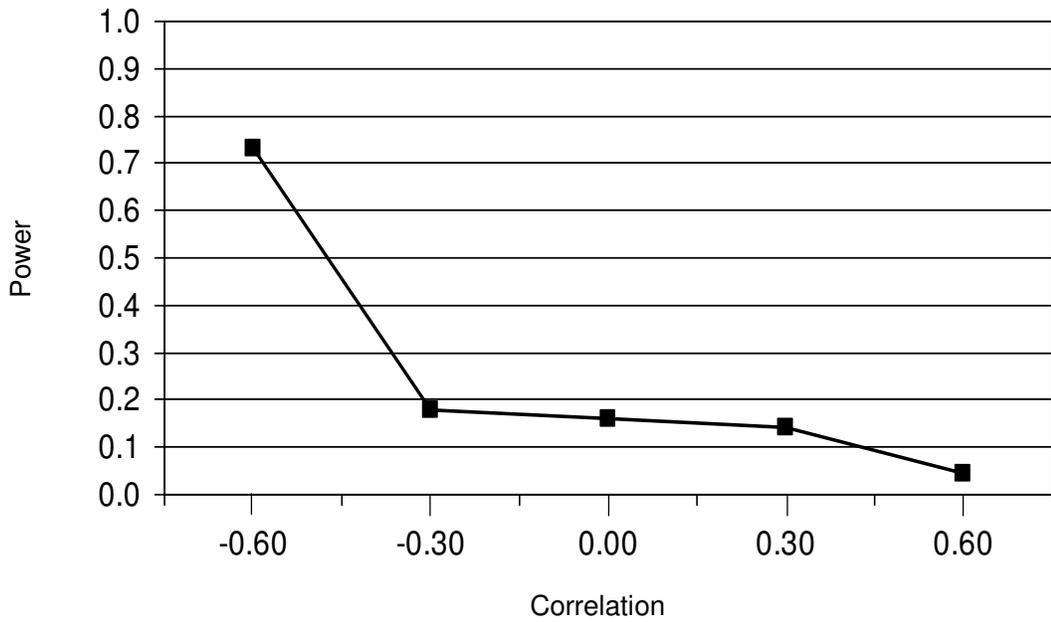
Figure 9. Adjusted power estimates for different within-class correlations between intercepts and slopes.
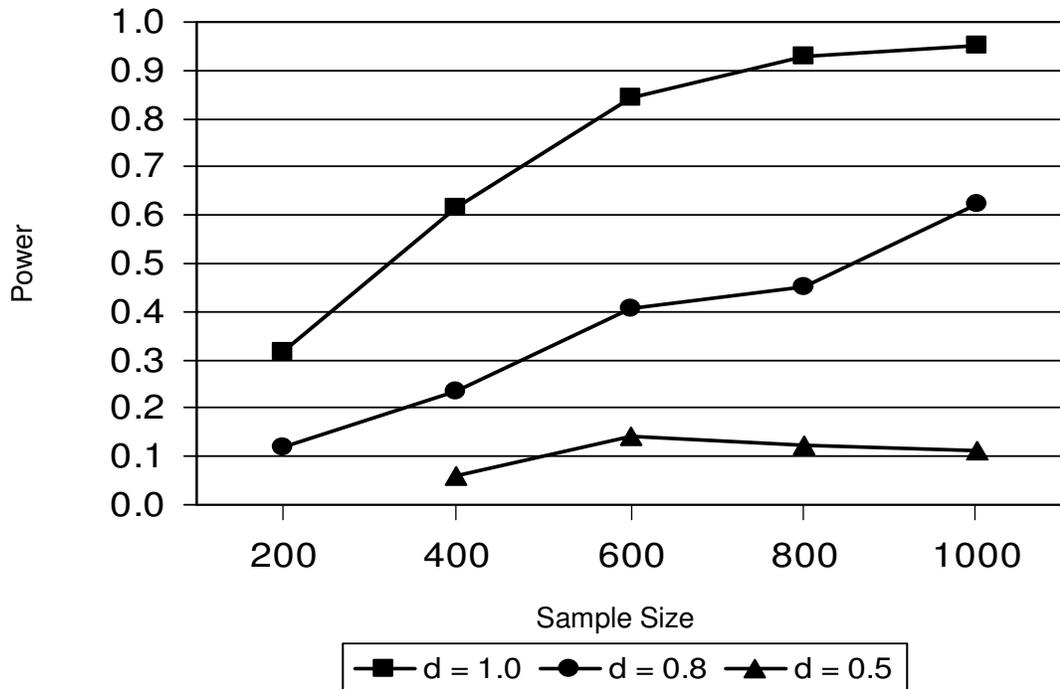
Figure 10. Adjusted power estimates for various sample sizes by standardized between-class differences in growth parameters (*d*s = 1.0, 0.8, and 0.5).
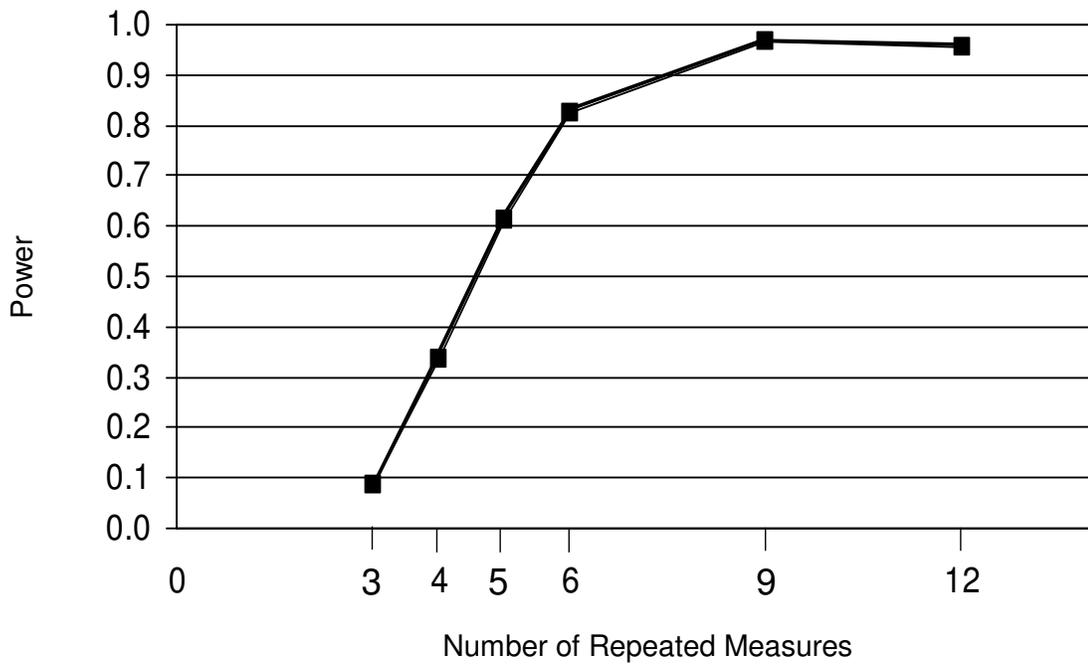
Figure 11. Adjusted power estimates by number of repeated measures.
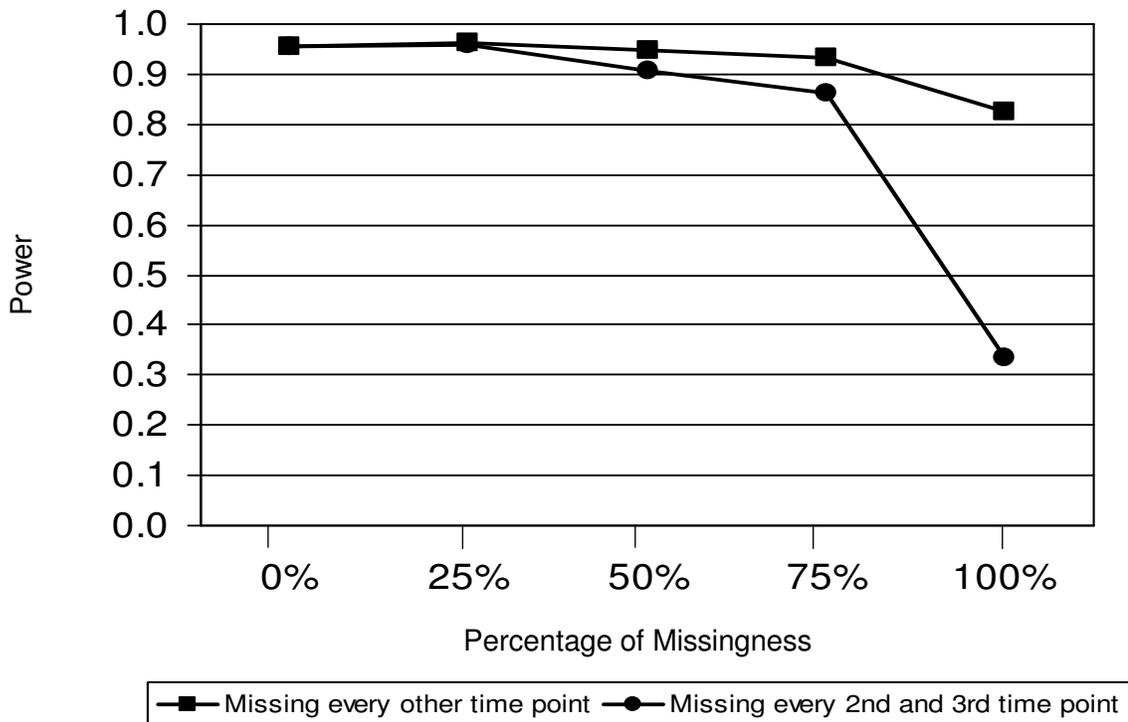
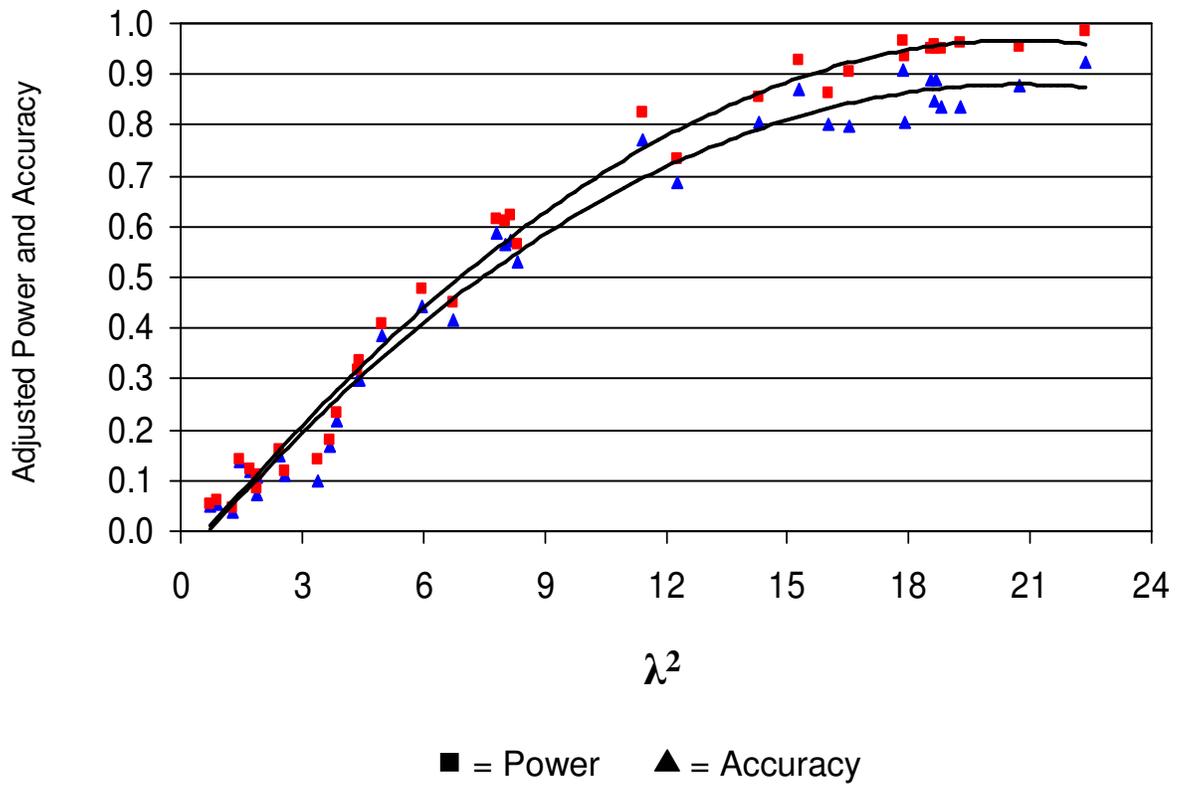*Figure 12.* Adjusted power estimates for two planned missingness designs, each with varying percentages of missing data.
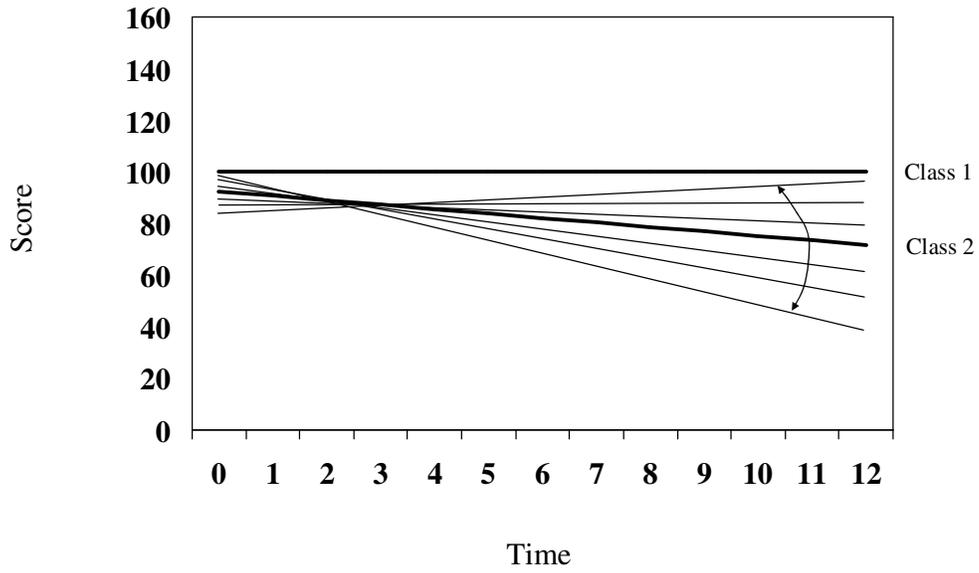
Figure 13. Estimates of adjusted power and accuracy by values of λ^2.

Negative Correlation Between Intercepts and Slopes

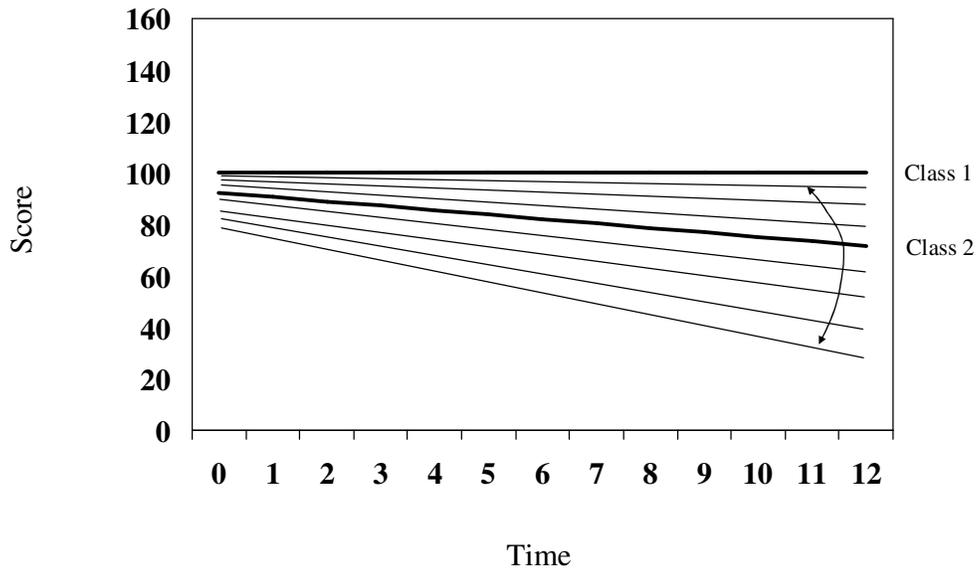## Positive Correlation Between Intercepts and Slopes

Figure 14. Growth trajectories for negative and positive correlations between intercepts and slopes (within Class 2).

145

About the Author

Eric Brown received dual Baccalaureate degrees in Economics and International Studies and a Master's degree in Economics from the University of South Florida. While in the Master's program, he worked as a data analyst at the Department of Child and Family Studies, Louis de la Parte Florida Mental Health Institute, University of South Florida, eventually accepting a position as a Research Assistant Professor. Currently, Dr. Brown works as a Research Associate at the Social Development Research Group at the University of Washington in Seattle.