Second-Generation Structural Equation Modeling with a Combination of Categorical and Continuous Latent Variables: New Opportunities for Latent Class/Latent Growth Modeling

Bengt Muthén
University of California, Los Angeles
Graduate School of Education & Information Studies
and
Department of Statistics *

Forthcoming in

New Methods for the Analysis of Change

A. Sayer & L. Collins (eds.)

Washington D.C.: APA

Second draft, November, 1998

*Invited presentation at the conference New Methods for the Analysis of Change, Pennsylvania

Abstract

New research provides an integration of categorical and continuous latent variable models. Given its generality, it is fitting to describe the emerging methodology as second-generation SEM, where the focus is on the generality of latent variable modeling (LVM). This LVM development promises to be extremely beneficial to growth modeling. The aim of this paper is to briefly introduce new LVM analyses in the form of General Growth Mixture Modeling (GGMM) and to show examples of the new analysis opportunities for growth modeling that are opened up. Five different GGMM examples are given representing five new types of growth analyses. The analyses are carried out by the new computer program Mplus (Muthén & Muthén, 1998a). The presentation is non-technical in order to reach applied researchers.

State University, October 11-13, 1998. This research was supported by grant K02 AA 00230 from NIAAA, by grant R21 AA10948 from NIAAA, and by grant 40859 from NIMH. The work has benefitted from discussions in Hendricks Brown's Prevention Science Methodology group and Muthén's Methods Mentoring Meeting. I thank David Francis for suggestions regarding the reading data application, Tom Harford for suggestions regarding the NLSY application, Hendricks Brown and Jason Liao for discussion related to intervention studies, Linda Muthén for useful comments, and Siek-Toon Khoo, Booil Jo, Noah Yang, and Christy Kim Boscardin for helpful research assistance. The email address is bmuthen@ucla.edu.

Key words: Growth curves, mixtures, latent trajectory classes

Complete address of author: Graduate School of Education & Information Studies, UCLA, Moore Hall, Box 951521, Los Angeles CA 90095-1521. Email address: bmuthen@ucla.edu

1 Introduction

This paper focuses on the interplay between growth curve modeling and structural equation modeling. Growth curve modeling is concerned with the study of individual differences in development over time, typically captured by random coefficients, i.e. growth parameters that vary across individuals. Structural equation modeling is concerned with relationships among observed and latent variables. The two areas connect due to the fact that random coefficients may be viewed as continuous latent variables (see, e.g. Meredith & Tisak, 1990). Growth modeling in a latent variable structural equation modeling (SEM) framework thereby benefits from the full generality of a such a framework. For example, in the SEM framework it is convenient to study regressions among the random coefficients, to do growth analysis in multiple populations, to analyze growth in latent variable constructs measured by multiple indicators, to analyze both parallel and sequential growth processes simultaneously, and to include other model parts that relate to the growth model such as mediators and distal outcomes. For an overview of some examples of this kind, see Muthén and Curran (1997).

The realization that growth curve modeling is possible using a latent variable conceptualization has enriched SEM research in recent years. New modeling epansions have otherwise been limited and methodological research has largely been concerned with refinements of estimators, standard errors, and tests of model fit. Growth modeling has enriched SEM, but by its general modeling framework SEM has also enriched growth modeling.

Latent variable structural equation modeling (SEM) is now about to take a radical leap forward in terms of modeling capabilities. As this paper describes, conventional SEM models using continuous latent variables are for the first time integrated with models using categorical latent variables. While categorical latent variable modeling has long traditions in the form of latent class analysis (LCA), latent profile analysis (LPA), and latent transition analysis (LTA), these traditions have been completely separated from SEM. Categorical latent variable modeling also encompasses finite mixture modeling in general and goes well beyond LCA, LPA, or LTA. New research provides an integration of categorical and continuous latent variable models. Given its generality, it is fitting to describe the emerging methodology as second-generation SEM, where the focus is on the generality of latent variable modeling (LVM). This LVM development promises to be extremely beneficial to growth modeling. The aim of this paper is to briefly introduce new LVM analyses in the form of General Growth Mixture Modeling (GGMM) and to show examples of the new analysis opportunities for growth modeling that are opened up. Five different GGMM examples will be given representing five new types of growth analyses. The presentation will be as non-technical as possible to reach applied researchers. The analyses of the five examples were carried out by the new computer program Mplus (Muthén & Muthén, 1998a). Input specifications are available from the author.

2 Conventional Latent Growth Modeling

Conventional growth modeling with random coefficients is a useful starting point for introducing the new modeling ideas. The basic ideas, scope, and limitations of conventional growth modeling are discussed here.

Figure 1 shows three ways of representing growth modeling: using a graph of individual trajectories; using a two-level model; and using a latent variable model. The top left panel of Figure 1 shows trajectories given by a linear growth model with hypothetical, but commonly seen parameter values. Individual lines are given for four individuals. The mean trajectory is also given as a solid line. It is seen that there is variation in the initial status at $t = 0$ as well as variation in the growth rate. Individuals starting higher grow faster.

The top right panel of Figure 1 translates the individual differences idea into a random coefficient model expressed as a multilevel model. The level-1 equation describes variation across time. Here, η_0 and η_1 represent intercept and slope coefficients in the regression of y on time t. The individual differences are captured by letting the intercept and slope vary across individuals, so that the η coefficients have subscript i. While this allows for across-individual heterogeneity, all individuals are assumed to come from one and the same population so that the analysis benefits from estimating all individuals together. The two level-2 equations describe variation across individuals for the intercepts and slopes and relates this variation to a background variable x, referred to as a time-invariant covariate (time-varying covariates can be included in the level-1 equa-

tion). The fact that the individuals come from a single population is reflected by the absence of a subscript i for the level-2 parameters of α and γ. Further single population parameters appear in the covariance matrix for the residuals ζ and for the residuals ϵ. Inserting the intercept and slope of the level-2 equations into the level-1 equation gives the model in a form referred to as the mixed linear model.

The bottom panel of Figure 1 translates the two-level model into a latent variable framework where the random coefficients are reconceptualized as latent variables. The level-1 equation and the level-2 equations correspond to a measurement and a structural part of a latent variable model, respectively. An advantage of growth modeling in a latent variable framework is that the framework opens up possibilities for general structural equation modeling, including regressions among random coefficients, multiple population analysis with flexible across-group differences in covariance matrix structures, growth modeling of latent variable constructs with multiple indicators, analysis of multiple processes, and the inclusion of other models parts such as mediational modeling and the analysis of distal outcomes.

2.1 A Simple Conventional Growth Model

Consider a simple linear growth model with an outcome y_{it} for individual i observed at timepoint t,

$$y_{it} = \eta_{0i} + \eta_{1i}\, x_t + \epsilon_{it}, \tag{1}$$

where η_{0i} is the intercept factor, η_{1i} is the growth rate factor, x_t represents a time-related score such as age, and ϵ_{it} is the time-specific residual assumed to be normally distributed. A given individual i has the values η_{0i} and η_{1i} on the two growth factors. The growth process for this individual develops over time as x_t changes as

$$\eta_{0i} + \eta_{1i}\, x_t. \tag{2}$$

This is individual i's trajectory, describing the systematic part of the variation of the outcome at different timepoints. The individual's outcome at a certain timepoint t, y_{it}, is equal to the sum of the systematic part of the variation plus the time-specific residual ϵ_{it}.

As mentioned above, the equation given in (1) is often referred to as the level-1 equation, describing the repeated measures over time. The level-2 equation describes the variation in the η_{0i} and η_{1i} factors as a function of covariates x,

$$\eta_{0i} = \alpha_0 + \sum_r \beta_{0r}\, x_{ri} + \zeta_{0i}, \tag{3}$$

$$\eta_{1i} = \alpha_1 + \sum_r \beta_{1r}\, x_{ri} + \zeta_{1i}, \tag{4}$$

where α coefficients are intercept parameters, β coefficients are regression weights for the covariates, and the ζ's represent residuals assumed to be bivariate normally distributed with covariance matrix $\mathbf{\Psi}$.

The parameters of (1), (3), and (4) can be estimated by maximum likelihood. Given such estimates, individual estimates of the growth factor values η_{0i} and η_{1i} can be obtained by the Bayesian approach of maximizing the posterior distribution of the factors

given the individual's observed data. In statistics this is referred to as Empirical Bayes estimation, while in psychometrics it is referred to as factor score estimation using the regression method.

2.2 Model Implications and Limitations

All of the five trajectories shown in Figure 1 follow the random coefficient growth model of (1), (3), and (4). The bold line is the average trajectory which is evaluated at the means of the two growth factors η_{0i} and η_{1i}, i.e. these factor means are inserted in (2). In addition, trajectories are shown for intercept growth factor values of one and two standard deviations below and above the mean, with corresponding slope factor values derived from the bivariate normal distribution. The variation across these five trajectories is due to the individual-specific influence of the covariate x_{ri} and the residuals ζ_{0i} and ζ_{1i}.

Some of the trajectories in Figure 1 represent quite different development over time. For example, if the outcome represents reading skills in grade 1, the two bottom lines labelled $i = 3$ and $i = 4$ may correspond to somewhat problematic or very problematic reading development, whereas the two top lines labelled $i = 2$ and $i = 1$ may correspond to good or excellent reading development. An assumption of the conventional growth model is that the data come from a single-population growth model that encompasses all these different types of trajectories, i.e. all individuals belong to one and the same population. In particular, it is assumed that the x covariates have the same

influence on the growth factors for all trajectories, i.e. the β covariate slopes of (3) and (4) are the same for all individuals. Using the example of the two types of good reading development versus the two types of poor reading development, this assumption of a homogeneous population may not be realistic. For example, the variation in reading development among poor readers may have a stronger influence from the school environment while the variation in reading development for good readers may have a stronger influence from the home environment, or vice versa. In addition, the variances for the residuals may differ across these two types of development, violating a second type of homogeneity assumption in the conventional growth model. The point is that although the conventional growth model captures individual differences in trajectories, it is not always realistic to assume that a single-population model can account for all types of individual differences.

More generally, the two types of development just discussed may have not only different antecedents but also different growth shapes, different concurrent processes, and different consequences. For example, problematic first-grade reading development may have a non-linear shape, may co-occur with the development of aggressive classroom behavior, and may increase the probability of subsequent deficits in achievement development or school dropout. This type of developmental heterogeneity is presumably quite common. In mental health, drug, and alcohol research the recognition of heterogeneity has led to theories of multiple developmental pathways (see, e.g. Moffitt, 1989 on adolescent-limited versus life-course-persistent antisocial behavior), subtypes (see, e.g. Zucker, 1994 and Schulenberg, O'Malley, Bachman, Wadsworth & Johnston, 1996 on

alcoholism and Nagin, Farrington, & Moffitt, 1995 on criminal offenders), and different disease processes (see, e.g. Pearson, Morrell, Landis, Carter, and Brant, 1994 on prostate cancer development). The Pearson et al. (1994) study illustrates some key modeling ideas. It examined the development of prostate specific antigen (PSA) and related this to later occurrence of tumors and cancers. While the normative development of PSA over age can be described by a linear growth model with variation in intercept and slope, development leading to tumors and cancers is characterized by a change from linear to exponential PSA growth, where the severity of the outcome corresponds to different forms of the growth curves.

In summary, the conventional growth model allows heterogeneity corresponding to different growth trajectories across individuals and captures that by variation in the continuous growth factors. However, the conventional growth model cannot capture heterogeneity that corresponds to qualitatively different development. While multiple-group SEM growth modeling is a flexible tool for studying qualitatively different development across individuals belonging to known groups, the problem here is that the group membership is typically not known but need to be inferred from the data. The next section discusses new methodology which introduces categorical latent variable modeling to greatly add to the capability of capturing heterogeneity in development.

3 Second-Generation Structural Equation Modeling

Second-generation structural equation modeling uses a combination of categorical and continuous latent variables. The ideas can be described with the help of the prototypical model structure of Figure 2. The figure shows four different modeling frameworks, labelled A, B, C, and D. These will be discussed in turn with an emphasis on C and D which provide the new growth modeling opportunities to be studied in the remainder of this paper. Four examples will be given of framework C analysis and one example will be given of framework D analysis.

3.1 A: Conventional Latent Variable Growth Modeling

The modeling framework labelled A has observed continuous outcomes or latent variable indicators, latent continuous variables, and observed background variables. This encompasses conventional SEM as it has been practiced in the last 25 years (see, e.g. Bollen, 1989). When the observed outcomes represent repeated measures over time, the latent variables are the growth factors and the background variables are the covariates as in (1), (3), and (4). Framework A was used in Muthén and Curran (1997) in their discussion of extensions of conventional growth modeling, particularly for randomized treatment-control studies.

3.2 B: Latent Class Modeling

The modeling framework labelled B has categorical latent class indicators, a categorical latent class variable, and observed background variables. This includes conventional latent class analysis (see, e.g. Clogg, 1995) as well as more recent extensions that add covariates (see, Dayton & McReady, 1988; Formann, 1992; van der Heijden, Dessens & Bockenholt, 1996; Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997). When the latent class indicators represent repeated measures over time, latent transition analysis (see, e.g. Collins & Wugalter, 1992) specifies a latent class variable for each timepoint and analyzes the transition probabilities between classes.

3.3 C: Finite Mixture Modeling

The modeling framework labelled C extends framework A by the categorical latent class variable. The broken arrow from the latent class variable to model framework A indicates that the parameters of A can be different for different latent classes. The analogy with conventional SEM is that the latent classes represent multiple populations or groups, but in contrast to multiple-group SEM the group membership is unobserved. This analogy makes it clear that framework C provides very flexible modeling given that multiple-group SEM allows group differences in any of the parameters. Model C analysis also produces a counterpart to factor scores for continuous latent variables. Given that the latent class variable is categorical, the factor score notion is generalized to posterior probabilities of membership in the different latent classes. A person may be classified

into the class that has the highest probability.

3.3.1 Mixture Models for Clustering

In statistical work, the latent classes are viewed as mixture components or missing data. A key reference for mixture analysis is Titterington, Smith and Makov (1985). Most of the statistical research using framework C does not include the latent continuous variables. A typical application is a multivariate model for the outcomes where the mixture components have the same covariance matrix and different mean vectors. This is a form of cluster analysis given that posterior probabilities of class membership are produced (see, e.g. McLachlan & Basford, 1988). An example is the classic 3-component analysis of the Fisher Iris data (see, e.g. Everitt & Hand, 1981). In psychometrics, related classic work has considered latent profile analysis (see, e.g. Gibson, 1959; Bartholomew, 1987), assuming a diagonal covariance matrix for each component.

3.3.2 CACE Modeling

Little and Yau (1998) give a novel example of a latent class variable viewed as missing data. This draws on Rubin's causal model, introducing the concept of Complier Average Causal Effect (CACE) estimation. Here, latent classes corresponding to those who participate versus do not participate in a randomized treatment program are used to give an assessment of treatment effects comparing participants in the treatment group with potential participants in the control group. For those randomized into the treatment

the latent class membership is observed. For individuals in the control group, however, category membership is missing, i.e. class membership is a latent categorical variable. This is a promising new technique that can also be used in growth modeling contexts.

3.3.3 Mixture SEM

Framework C has been considered in an SEM context. For example, Blåfield (1980), Yung (1997), Jedidi, Ramaswamy, DeSarbo and Wedel (1996), and Jedidi, Jagpal and DeSarbo (1997) studied mixture confirmatory factor analysis and structural equation modeling. The models studied by these authors, however, are more limited than what is shown as framework C in Figure 2. This is because the models do not allow for the regression of the latent class variable on the background variables. This limitation is shared by most mixture modeling, two recent exceptions being Little and Yau (1998) and Nagin and Tremblay (1998). Not including the regression of the latent class variable on the background variables has two drawbacks. First, the probabilities of latent class membership are taken to be the same for individuals with different values on the background variables which may be unrealistic in many settings. Second, a two-stage estimation is necessary when exploring the characteristics of the latent classes. Instead of a joint maximum-likelihood analysis of the full model for the latent variable indicators, the latent continuous variables, the latent class variable, and the background variables, a second-stage analysis is needed where estimated posterior probabilities for the latent classes are related to the background variables.

3.3.4 Growth Mixture Modeling

The primary focus for framework C in this paper is growth mixture modeling. Verbeke and LeSaffre (1996) discussed a mixed linear model approach to random coefficient growth modeling with a mixture that allows for different means of the random coefficients. Extending the work of Nagin and Land (1993), Nagin and Tremblay (1998), Jones, Nagin, and Roeder (1998), and Roeder, Lynch, and Nagin (1998) discuss modeling of different trajectory classes with normal and non-normal outcomes. The Nagin et al. work also recognizes the importance of estimating the effects of class membership on covariates. However, for each class their modeling considers a fixed-effect growth model, not a random coefficient growth model. In this paper we will consider a flexible random coefficient growth mixture model based on framework C and, further on, framework D. Framework C may be viewed as providing model framework A for each latent class category and therefore allows for considerable modeling generality. The discussion based on framework C will center around four types of growth mixture examples. The discussion draws on recent work in Muthén, Brown, Khoo, Yang, and Jo (1998) and Muthén and Shedden (1998).

The first example shows how the growth mixture model provides a solution to the problem of heterogeneity in covariate influence discussed in the previous section. In terms of Figure 2, the arrow from the latent class variable can be seen as representing latent class membership influence on the arrow from the background variables to the latent continuous variables. This means that different latent classes have different slopes

in the regression of the latent continuous variables on the background variables.

The second example shows how the growth mixture model provides a solution to the problem of heterogeneity of growth curve shapes. Here, the arrow from the latent class variable can be seen as representing latent class membership influence on the means (intercepts) of the latent continuous growth factors. Different latent classes have different growth shapes because the growth factor means are different. Different latent classes may also have different number of growth factors; e.g. one class has linear growth and one class has quadratic growth.

The third example also concerns an analysis of growth curve shapes. While the second example is an exploratory analysis, the third example concerns confirmatory analysis with prespecified growth curve shapes that are of particular interest to the researcher.

The fourth example shows how the growth mixture model provides a solution to yet another form of growth shape heterogeneity. Here, the arrow from the latent class variable can be seen as representing latent class membership influence on the slopes (loadings) for the outcomes, i.e. the latent variable indicators, on the latent continuous variables. These slopes capture the time-related scores x_t in (1) and reflect the curvature of the trajectory. The particular application in this case is piecewise modeling where different stages of development are captured by a growth model with different pieces corresponding to different growth rates. Conventional piecewise modeling has the weakness of assuming that every sample unit makes the transition from one piece

to the next at the same, known timepoint. In contrast, mixture modeling allows for individually-varying transition points between the different stages of development.

3.4 D: General Latent Variable Modeling (LVM)

The modelling framework labelled D is a combination of the other frameworks and will be referred to as general latent variable modeling, or LVM for short. This may be seen as second-generation structural equation modeling given that the modeling capabilities are vastly enhanced relative to conventional SEM. Because the generality is gained by the introduction of both categorical and continuous latent variables, it seem appropriate to focus on the concept of latent variables and use the term LVM. LVM expands latent class modeling, finite mixture modeling, and SEM. On the one hand LVM recognizes the usefulness of expanding the latent class framework. Latent class analysis concerns a particular mixture model of independence among a set of categorical latent class indicators. In contrast, LVM lets the latent classes also influence more general mixture models for the other model parts of framework A. On the other hand LVM recognizes the usefulness of expanding finite mixture modeling to include direct indicators of the mixture components. Finite mixture modeling infers mixture component membership from the distribution of the latent variable indicators and the background variables. In contrast, LVM can let the latent class and latent class indicator part of the model be a specific measurement model from which class membership can be inferred beyond the information from the observed variables in framework A.

Model framework D provides the fifth and final example. The fifth example uses the growth mixture model of the second example where the latent class variable determines a set of growth shapes. A new type of application is obtained by letting latent trajectory class membership corresponding to these growth shapes influence the probability of a later outcome as represented by a single latent class indicator.

Technical aspects of LVM and its estimation by maximum likelihood is described in Muthén, Shedden, and Spisic (1998) and will not be discussed here. The analyses of all five examples presented here were carried out by this method as implemented in the computer program Mplus (Muthén & Muthén, 1998a). Mplus is a new program that replaces the author's LISCOMP program. Despite its ability to carry out such general types of models as indicated by framework D, Mplus has a simpe model specification language without matrices or equations that is suitable for applied researchers.

3.4.1 General Growth Mixture Modeling

From a growth modeling perspective, the new growth modeling opportunities made available in LVM may be labelled General Growth Mixture Modeling (GGMM) as was done in Muthén et al. (1998). In summary, GGMM goes beyond conventional random coefficient growth modeling by using latent trajectory classes which allow for heterogeneity with respect to the influence of antecedents, growth shapes, concurrent outcomes, and later consequences. Two additional features of the analysis are particularly noteworthy, confirmatory analysis and estimation of class membership probabilities.

Confirmatory analysis is a key element in conventional SEM. A priori hypotheses can be captured by parameter restrictions. For instance, factor loadings may be fixed at zero to reflect the belief that a certain factor is not measured by a certain indicator. GGMM in framework D offers the same kind of confirmatory analysis, but also a second type of confirmatory analysis which does not concern hypotheses about parameters but about individuals' class membership.

Confirmatory analysis with respect to parameters is very useful in growth settings. For example, there may be quadratic growth for one class, but the growth shape of a second class is hypothesized to be linear. Here, the quadratic growth factor mean and variance are fixed to zero for the second class. Or, one class may have the mean of one of its growth factors constrained to be larger than that of another class. The mixture analysis benefits greatly from these types of confirmatory restrictions. In contrast, conventional finite mixture analysis is typically exploratory. Finite mixture analysis is known to sometimes give rise to numerical analysis problems, such as nonconvergence and multiple maxima of the likelihood. Confirmatory mixture analysis limits the occurrence of such problems. Recent experiences show that growth mixture analysis is a relatively well-behaved form of mixture analysis given the multivariate trend information in the data, as well as the possibilities for formulating confirmatory hypotheses.

Confirmatory analysis with respect to the class membership of the individuals is a feature not found in SEM. With a latent class variable, however, a researcher may want to incorporate the hypothesis that certain individuals are known to represent typical trajectories corresponding to a certain class. This knowledge may be due to auxiliary

information or because the individual displays such a typical growth pattern. For example, students who are known to have been diagnosed as having reading disability in second grade may have a characteristic reading skills trajectory in first grade. Individuals with known class membership are sometimes referred to as training data. As is the case with using parameter restrictions, the numerical performance of the mixture analysis benefits greatly from incorporating training data. It may be noted that multiple-group SEM corresponds to the case of all sample units contributing training data so that the latent class variable is in effect an observed categorical variable.

GGMM is a promising tool for analysis of randomized trials. A discussion of GGMM in preventive intervention settings is given in Muthén et al. (1988). Interventions often show an interaction between treatment and characteristics of the individual. Important individual characteristics can be captured by developmental trajectory classes. Differences between classes may refer to baseline characteristics, such as the initial status of a growth model (Muthén & Curran, 1997), but also growth rate and more generally growth shape. GGMM analysis allows the effect of treatment to vary across trajectory class and is therefore able to give a more detailed assessment of treatment effects. It is also useful to add the notion of partial compliance and use Complier Average Causal Effect (CACE) estimation based on latent compliance classes. The use of GGMM in conjunction with CACE gives a flexible analysis framework for randomized trials; for some initial work see Muthén and Jo (1998).

A final general comment on GGMM concerns the benefit of estimating an individual's posterior probabilities of class membership. This is the counterpart to factor scores

for conventional SEM. Given an estimated model, each individual obtains a posterior probability estimate for each class computed as a function of the model parameter estimates and the individual's values on his or her observed variables. The class to which the individual most likely belongs can therefore be determined.

Estimation of class membership probabilities is of particular interest with longitudinal data. Consider as an example a model for development of reading skills through first and second grade. A GGMM analysis estimates the parameters for this development. The class membership probabilities can then be estimated for a new student before the student reaches the end of second grade using only the subset of the repeated measures available at that point in time. It is of interest to study with which precision such probabilities can be estimated at different points in time. This allows for a mechanism for early prediction of problematic development. The analysis approach may be particularly useful in diagnostic or preventive intervention settings. Not only may individuals belonging to different classes benefit differently from a treatment program but the analysis could possibly also guide in the choice of treatment variation.

4 GGMM Examples

4.1 Example 1. Analysis of Math Achievement Development: Differences in Influence of Covariates

Example 1 uses a subset of the data from the Longitudinal Study of American Youth (LSAY). The analyses concern females in the younger cohort measured at four time-points, grade 7 - grade 10, beginning in 1987. Math achievement items and background information from parents on mother's education and home resources are used. Conventional random coefficient growth analysis of both females and males indicate that mother's education and home resources are both predictive of the initial status in grade 7 as well as the growth rate over grades. The residual variances are significant for both initial status and growth rate.

In example 1, a 2-class growth model is explored for 984 females with complete data on the variables. First, the analysis is carried out without the two covariates (unconditional analysis) and next with the two covariates (conditional analysis). Figure 3 shows the general GGMM framework applied to this example. The latent categorical variable with two classes is represented by the circle labelled c. The intercept factor is labelled I and the slope factor is labelled S. The analysis is first carried out without the two covariates mother's education and home resources (unconditional analysis) and then including these covariates (conditional analysis). The prediction of latent class membership from the covariates will be introduced as the last step of the analysis.

4.1.1 Unconditional Analysis

A baseline model lets the growth factor means for the intercept factor and the slope factor vary across the two latent classes while all other parameters are held invariant across the classes. This results in a log likelihood value of $-12,692.249$ which should be compared to the conventional single-class growth model with log likelihood value of $-12,729.608$ with three fewer parameters (two growth factor mean parameters and one class probability parameter).

Comparison of models with different number of classes can be accomplished by using standard information criteria such as AIC and BIC. The measures are based on the negative of the log likelihood of the model, with an increasing penalty for an increasing number of parameters. Smaller values indicate better models. The BIC penalty is the number of parameters r multiplied by $ln\ n$, where n is the sample size (Schwartz, 1978),

$$BIC = -2\ logL + r\ ln\ n. \tag{5}$$

The BIC values for the 1- and 2-class models are $25,521.24$ and $25,467.20$, respectively. The BIC criterion favors the 2-class model over the 1-class model in this example. More than two classes may be suitable for these data, but such models will not be explored here.

The 2-class baseline model imposes class-invariance of the covariance matrix for the two growth factors as well as class-invariance of the variances for the residuals of the outcomes. Letting the variances for the residuals of the outcomes vary across the two classes gives a log likelihood value of $-12,606.93$, corresponding to a likelihood-ratio χ^2

difference test value with 4 degrees of freedom equal to 170.6. This indicates that the residual variances need to be different across classes.

Also letting the factor covariance matrix differ across the two classes gives a log likelihood value of $-12,595.24$ and $\chi^2(3) = 23.4$, indicating a need for these parameters to also be different across the classes. For this model there are about 43% in class 1, where compared to class 2, class 1 is characterized by low starters who grow slowly. For class 1 the outcome residual variances are estimated as $16.30, 16.29, 22.10$, and 49.4, while for class 2 they are estimated as $12.63, 11.26, 8.15$, and 8.12. The largest difference is for grade 10, showing a larger degree of time-specific variance for the low-performing class. For class 1 the intercept variance is 53.76 while for class 2 the intercept variance is 39.85. For class 1 the slope variance is 1.76, while for class 2 it is 0.17. All of these variances are significant except the last one mentioned. While the growth rate variation for class 1 is significant, the insignificant slope variance for class 2 indicates that students in this well-performing class develop equally fast. In summary, the two classes show quite different development.

4.1.2 Conditional Analysis

The conditional analysis attempts to capture some of the variation in the growth factors using the two covariates mother's education and home resources. In this analysis, the slope factor variance is fixed at zero for class 2 based on the unconditional analysis (an analysis letting this variance be free did not show a significant influence of the covariates

on the growth rate, nor a significant residual variance). For both class 1 and class 2 mother's education and home resources were significant predictors of the intercept factor, i.e. the initial status at grade 7. For class 1, home resources but not mother's education was a significant predictor of the growth rate. In this way, the 2-class analysis shows that home resources is an important factor for math achievement growth, but only for those not developing well. In contrast, the conventional analysis mistakenly concludes that home resources is important for growth for all individuals. With respect to the covariates, the conventional single-class analysis results are more in line with the class 1 results than the class 2 results. This indicates that the results for the majority class do not always determine the outcome of a conventional analysis. The 2-class analysis is of educational importance in that it carefully delineates the impact of home resources on math achievement growth, showing that home resources is important only for a subset of individuals.

As shown in the GGMM framework C in Figure 3, the latent class membership may also be related to the two covariates. As part of the GGMM analysis, the logit regression of c on mother's education and home resources showed that home resources had a significant influence on the probability of membership in class 1.

4.2 Example 2. Analysis of Normative and Non-Normative Development in Heavy Drinking: Exploring Growth Curve Shapes

Example 2 uses a subset of data from the National Longitudinal survey of Youth (NLSY). The outcome variable is frequency of heavy drinking (6 or more drinks on one occasion) during the last 30 days. Covariates are gender, ethnicity, family history of alcohol problems, early start of regular drinking (age 14 or less), and dropping out of high school. For this illustration heavy drinking is considered for one of eight NLSY cohorts covering ages 18, 19, 20, 24, and 25. The sample size is 934.

A quadratic growth model has been found suitable for the development of heavy drinking over ages 18 to 37 (Muthén & Muthén, 1988b). The growth curve shape at the average of the three growth factors is given as the bottom solid curve in Figure 4, showing an increase from 18 to 21 with a subsequent decrease. This normative growth curve shape has also beeen found for delinquent behavior and illicit drug sampling. One may ask if there are also other, non-normative, growth curve shapes represented in this population. For example, some individuals may show an increase from 18 to 21 but no downturn as illustrated by the solid line. Given that the conventional, single-class, quadratic random coefficient model allows for individual variation in all three growth factors, the solid line trajectory shape is actually included as a special case in the conventional model. The question is if a better fit to the data can be obtained by a 2-class growth model where the classes differ in their means on the three growth factors. In line with the GGMM model

framework C, such a 2-class model can be used to estimate the influence of covariates on the probability of class membership. The model can also be used to estimate each individual's most likely trajectory class membership.

GGMM analyses on the NLSY data were performed in Muthén and Shedden (1998). A 2-class model resulted in two different solutions, where the two solutions shared the normative curve for the majority, but had different non-normative curves. A 3-class solution showed all three kinds of curve shapes and was also found to fit the data better in terms of BIC values. The Muthén and Shedden (1998) 3-class solution is shown in Figure 5. The normative class probability was estimated as 81.0%, whereas the two non-normative class probabilities were estimated as 7.1% and 11.9%, respectively.

As part of the GGMM analysis, membership in the three classes is also related to the set of covariates. Figure 6 shows the estimates of the corresponding multinomial logistic regression. Here, High refers to the class in Figure 5 that is high already at age 18, Increase refers to the class that increases over age, and Norm refers to the normative class. In the column labelled High versus Norm, the coefficients show how the odds of belonging to the high class compared to the normative class is significantly (bolded numbers) increased for individuals who are male and have early onset, and is significantly decreased for individuals who are black and did not go to college. In the columns labelled Increase versus Norm, the coefficients show how the odds of belonging to the increasing class compared to the normative class is significantly increased for individuals who are male and who dropped out of high school.

4.3 Example 3. Analysis of Reading Skills Development: Confirmatory Analysis of Growth Curve Shapes

Example 3 uses a subset of data from the study Early Assessment of Reading Skills (EARS). Here, it was of interest to study reading skills development in first grade and relate this to Kindergarten precursor skills to see if Kindergarten growth trajectory patterns were predictive of first grade development. In this analysis, four repeated measures of first-grade word recognition will be related to four repeated measures of Kindergarten phonemic awareness skills for a subsample of 410 children.

As a first attempt to relate first-grade development to Kindergarten precursors, Muthén, Khoo, Francis, and Kim Boscardin (1998) considered conventional growth modeling in a latent variable framework in line with Figure 7. Recognizing that the first-grade outcome of word recognition and the Kindergarten phonemic awareness are in different metric, two growth sequential growth models were specified and related to each other. For each, the intercept factor was defined as the exit level rather than the starting point. As seen in Figure 7, the intercept factor I_w defined for word recognition at exit from first grade is related to the Kindergarten phonemic awareness exit level I_p and slope S_p. Analysis of this model, showed a significant influence of I_p on I_w but no significant influence of S_p on I_w. The lack of significant influence of the phonemic awareness slope S_p was a surprise and it was decided to make a more thorough investigation of this using GGGMM.

The GGMM analysis took the conventional growth model in Figure 7 as a starting

point to define a set of trajectory classes types of particular interest. For the Kindergarten process, three different levels of I_p were identified, the mean of the I_p factor, and plus and minus one standard deviation away from the mean. For each of these three Kindergarten exit levels two levels of S_p slope values were chosen, plus and minus one half standard deviation away from the conditional mean of S_p given I_p. The resulting six classes are shown in the left panel of Figure 8. In the analysis, the six classes are specified by fixing the two Kindergarten growth factor means at the values just mentioned. The childrens' class memberships are, however, not fixed but estimated in the analysis. Furthermore, the analysis estimates the means of the two first-grade growth factors for each of the six classes. This addresses the key aspect of the analysis: what first-grade trajectories are found for children in different Kindergarten trajectory classes?

A particularly interesting class comparison concerns class 5 and class 6. These classes have the same Kindergarten exit level but differ in their Kindergarten growth rate. One may ask which class will perform better in first grade. Class 5 has the advantage of having been at this exit level longer than class 6, but class 6 shows quicker learning. The GGMM analysis results given in the right panel of Figure 8 show that class 6 performs considerably better in first grade than class 5. This means that the S_p slope is important for I_w, contrary to the conventional analysis result. A higher Kindergarten slope is, however, not always advantageous as comparisons between class 3 and class 4 and between class 1 and class 2 show. For these two higher Kindergarten exit levels a lower S_p slope is more advantageous. In this way, the interaction between I_p and S_p is not the same at different I_p levels. Even if possible, this heterogeneous interaction precludes

an analysis with a model of interactions among continuous latent growth factors. Also, the heterogeneous interaction may explain why the conventional analysis found S_p to be an insignificant predictor of I_w - at some Kindergarten exit levels a high S_p value is advantageous and some other levels it is not. The GGMM analysis creates a framework for early prediction of class membership based on Kindergarten performance.

4.4 Example 4. Piecewise Growth Modeling with Individually-Varying Transition Points

Example 4 concerns heterogeneity in the relationship between the continuous latent variable indicators and the continuous latent variables, i.e. the measurement relations. Figure 9 gives such an example. In Figure 9, the class 1 trajectory follows a conventional piecewise growth model where the first piece shows linear growth up to $t = 3$, followed by a second piece of linear growth with a steeper slope. Conventional latent growth modeling analyzes such data with a model described in the path diagram below, where the two slope factors have different means and variances.

A weakness of piecewise growth modeling is that the point of transition from one stage to the next needs to be both known and the same for all individuals in the sample. GGMM allows more flexible piecewise growth modeling where different unknown classes of individuals make the transition at different timepoints. Figure 9 shows an example with two different transition points, at $t = 3$ and at $t = 4$. The two classes have the same number of outcome variables and the same number of growth factors, but the growth

factor "loadings" are different for the two classes. Simulation studies of the model given in Figure 9 show that the 2-class model parameters of this extended piecewise growth model are easily recovered. The model would seem to be of particular interest in intervention studies where a certain treatment may take effect after different amounts of time for different classes of individuals.

4.5 Example 5. Analysis of Normative and Non-Normative Development in Heavy Drinking: Growth Curve Shapes as Predictors of Distal Outcomes

Example 5 returns to the Muthén-Shedden NLSY heavy drinking example discussed in Example 2. Here, the 3-class model is embedded in a larger model shown in Figure 10. While the heavy drinking concerns behavior from ages 18 to 25, the model will now be used to predict a binary variable of alcohol dependence at age 30. As in Example 2, the latent trajectory class variable in Figure 10 is seen to influence the growth curve shape of heavy drinking development via the growth factor means and the latent trajectory class variable is predicted by covariates. The additional feature in Figure 10 is the prediction of alcohol dependence by trajectory class. This is a logistic regression relationship, although the predictor variable is latent. In Figure 2 terms, alcohol dependence is an example of a latent class indicator variable and the model an example of model framework D.

The use of a latent trajectory class variable as a predictor solves a dilemma that

occurs when trying to use the continuous growth factors as predictors. Consider as an example the two trajectory classes of Figure 4. Here the growth factor representing linear growth rate concerns the rate of increase right after age 18. This linear growth rate factor is not, however, a suitable predictor of later problematic alcohol use outcomes because both classes have a high positive slope value. Likewise, if there is a third class which is high already at age 18, a non-positive slope may be predictive of problematic outcomes. The key issue is that the growth factor values interact in determining the growth shape and it is the shape that is predictive.

Figure 11 shows the resulting estimates of the relation between trajectory class and alcohol dependence. While the normative class ($c = 3$) has an estimated probability of 0.08 of developing alcohol dependence at age 30, the two non-normative classes have elevated probabilities of 0.36 ($c = 2$) and 0.24 ($c = 1$), respectively. Given that the probabilities vary so strongly as a function of the latent trajectory class membership, it is of great importance to be able to predict the class membership as early as possible using both covariate information and early information on heavy drinking development.

5 Conclusions

It is clear from the five examples that General Growth Mixture Modeling (GGMM) is an important new development for the study of change. The new modeling opportunities will enrich growth modeling and be better able to reflect more complex ideas of development based on substantive theories in various fields. GGMM is further strengthened by

fitting into a general latent variable framework with which it can be combined. In this way, it is likely that growth modeling and latent variable modeling will both continue to benefit from developments in the other area.

References

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

Bartholomew, D.J. (1987). Latent variable models and factor analysis. New York: Oxford University Press.

Blåfield, E. (1980). Clustering of observations from finite mixtures with structural information. Unpublished doctoral dissertation, Jyväskyla studies in computer science, economics, and statistics, Jyväskyla, Finland.

Bollen, K.A. 1989. Structural equations with latent variables. New York: John Wiley.

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.

Collins, L.M. & Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. Multivariate Behavioral Research, 27, 131-157.

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

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

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. Journal of the American Statistical Association, 87, 476-486.

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

Jedidi, K., Jagpal, H. S. & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. Marketing Science, 16, 39-59.

Jedidi, K., Ramaswamy, V., DeSarbo, W. S. & Wedel, M. (1996). On estimating finite mixtures of multivariate regresson and simultaneous equation models. Structural Equation Modeling, 3, 266-289.

Jones, B.L., Nagin, D.S. & Roeder, K. (1998). A SAS procedure based on mixture models for estimating developmental trajectories.

Little, R.J. & Yau, L.H.Y. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. Psychological Methods, 3, 147-159.

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

Moffitt, T.E. (1993). Adolescence-limited and life-course persistent antisocial behavior. Psychological Review, 100, 674-701.

Muthén, B. & Curran, P. (1997). General longitudinal modeling of individual differ-

ences in experimental designs: a latent variable framework for analysis and power estimation. Psychological Methods, 2, 371-402.

Muthén, B. & Jo, B. (1998). Growth modeling in randomized trials with latent trajectory classes and non-participation: an overview of new analysis opportunities.

Muthén, B., Khoo, S.T., Francis, D., & Kim Boscardin, C. (1998). Analysis of reading skills development from Kindergarten through first grade: an application of growth mixture modeling to sequential processes.

Muthén, L.K. & Muthén, B. (1998a). Mplus user's guide. Los Angeles: Muthén & Muthén.

Muthén, B. & Muthén, L. (1998b). The development of heavy drinking from age 18 to 37 in a U.S. national sample. Under review in Journal of Studies on Alcohol.

Muthén, B. & Shedden, K. (1998). Finite mixture modeling with mixture outcomes using the EM algorithm. Forthcoming in Biometrics.

Muthén, B., Shedden, K. & Spisic, D. (1998). General latent variable mixture modeling.

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

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

Nagin, D.S. & Tremblay, R.E. (1998). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and non violent juvenile delinquency. Forthcoming in Child Development.

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

Pearson, J.D., Morrell, C.H., Landis, P.K., Carter, H.B., & Brant, L.J. (1994). Mixed-effect regression models for studying the natural history of prostate disease. Statistics in Medicine.

Roeder, K., Lynch, K.G., & Nagin, D.S. (1998). Modeling uncertainty in latent class membership: A case study in criminology.

Schulenberg, J., OMalley, P.M., Bachman, J.G., Wadsworth, K.N., & Johnston, L.D. (1996). Getting drunk and growing up: Trajectories of frequent binge drinking during the transition to young adulthood. Journal of Studies on Alcohol, May, 289-304.

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

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

van der Heijden, P.G.M., Dessens, J. & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. Journal of Educational and

Behavioral Statistics, 21, 215-229.

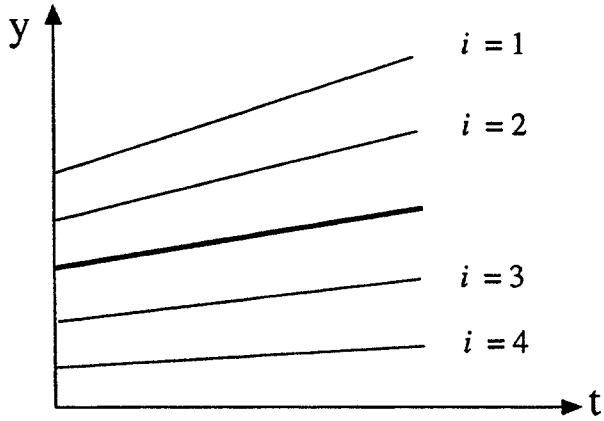
Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association, 91, 217-221.

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

Zucker, R.A. (1994). Pathways to alcohol problems and alcoholism: A developmental account of the evidence for multiple alcoholisms and for contextual contributions to risk. In: R.A. Zucker, J. Howard & G.M. Boyd (Eds.), The development of alcohol problems: Exploring the biopsychosocial matrix of risk (pp. 255-289) (NIAAA Research Monograph No. 26). Rockville, MD: U.S. Department of Health and Human Services.

Figure 1.

Three representations of individual differences in growth.



Individual i at time t:

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

$$(2a)\quad \eta_{0i} = \alpha_0 + \gamma_0 X_i + \zeta_{0i}$$

$$(2b)\quad \eta_{1i} = \alpha_1 + \gamma_1 X_i + \zeta_{1i}$$

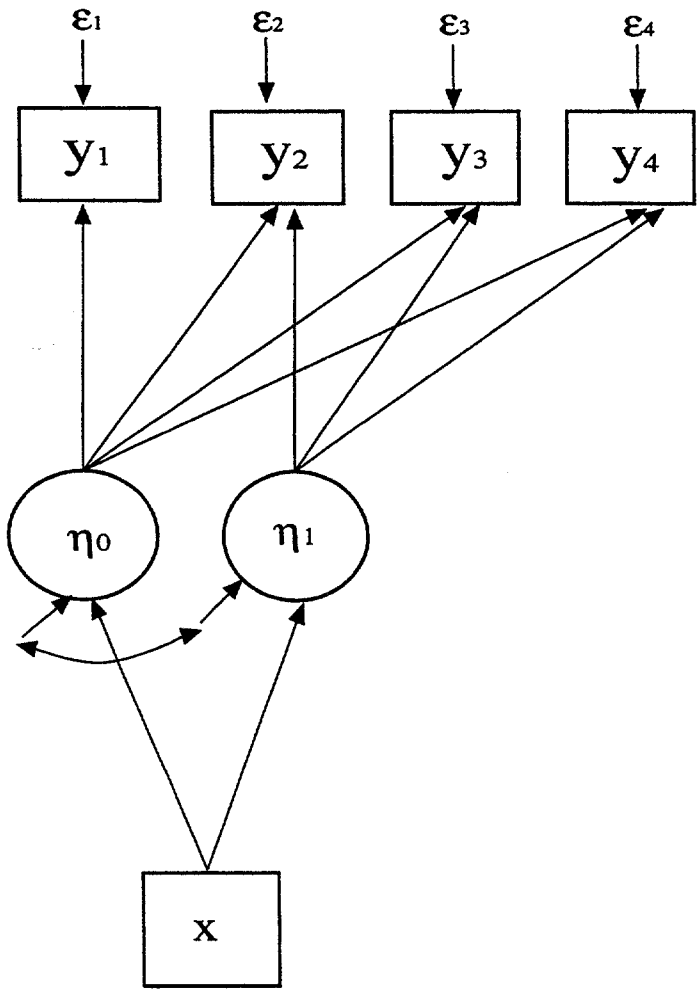


Figure 2.

A general framework for latent variable modeling.

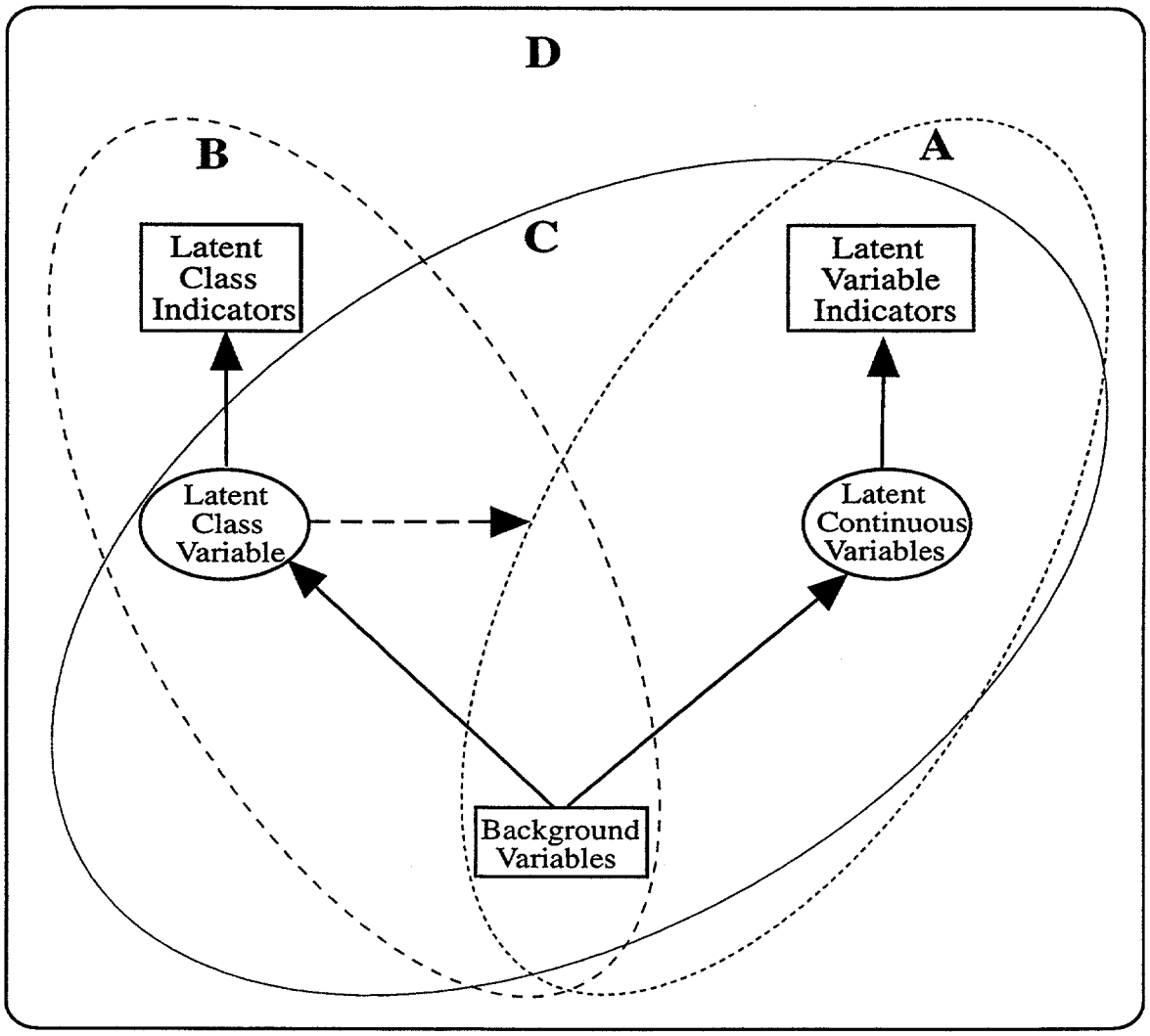

D
B
A
C
Latent Class Indicators
Latent Variable Indicators
Latent Class Variable
Latent Continuous Variables
Background Variables

Figure 3.

Math achievement modeling.

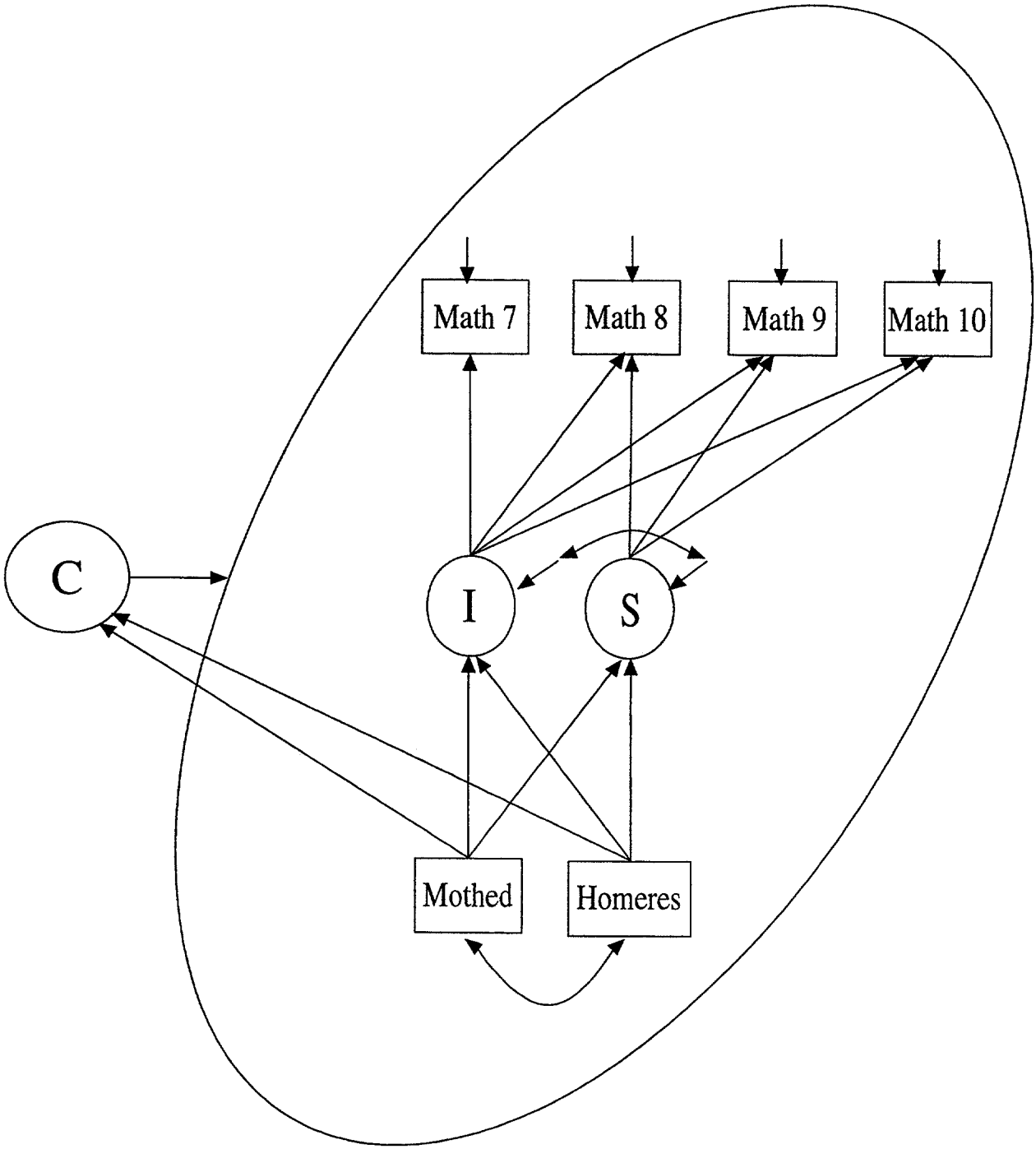

Math 7
Math 8
Math 9
Math 10
C
I
S
Mothed
Homeres

Figure 4.

Modeling of heavy drinking.

Example: NLSY Heavy Drinking
Two Latent Trajectory Classes

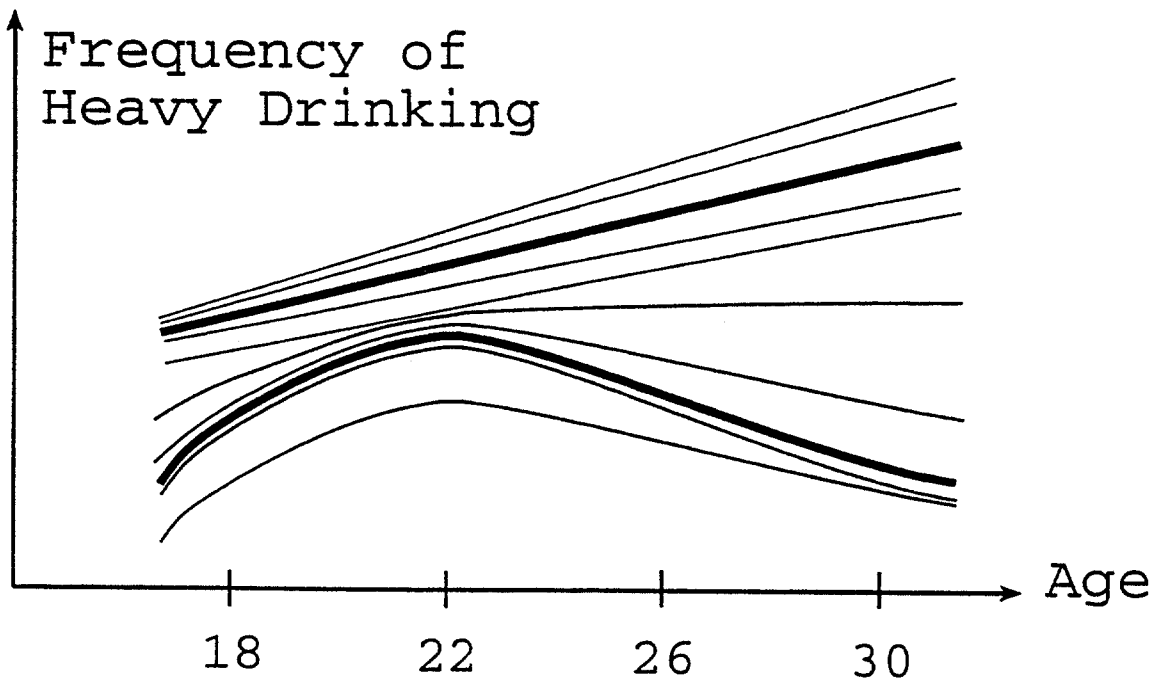


Figure 5.

Estimated heavy drinking model: curves.

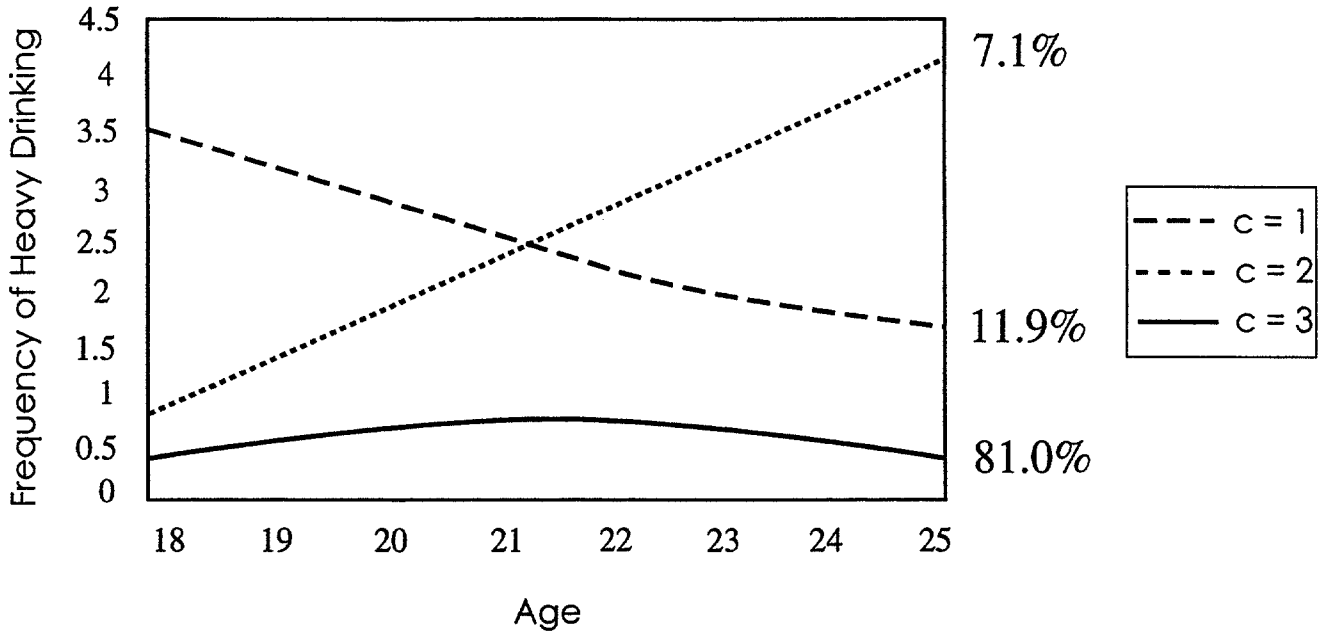

Frequency of Heavy Drinking
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
18
19
20
21
22
23
24
25
Age
7.1%
11.9%
81.0%
c = 1
c = 2
c = 3

Figure 6.

Estimated heavy drinking model: covariates.

Predicting Trajectory Class Membership

Estimated Logit Coefficients:

Covariate (x)	High vs Norm	Increase vs Norm
Male	**1.25**	**1.48**
Black	**-1.60**	-.67
Hispanic	-.22	.74
Early Onset	**1.07**	.62
FH123	.62	.68
Dropout	.22	**.80**
College	**-.61**	-.04

Figure 7.

Reading skills development: conventional model.

Reading Precursors in Kindergarten

Growth Modeling of Sequential Processes

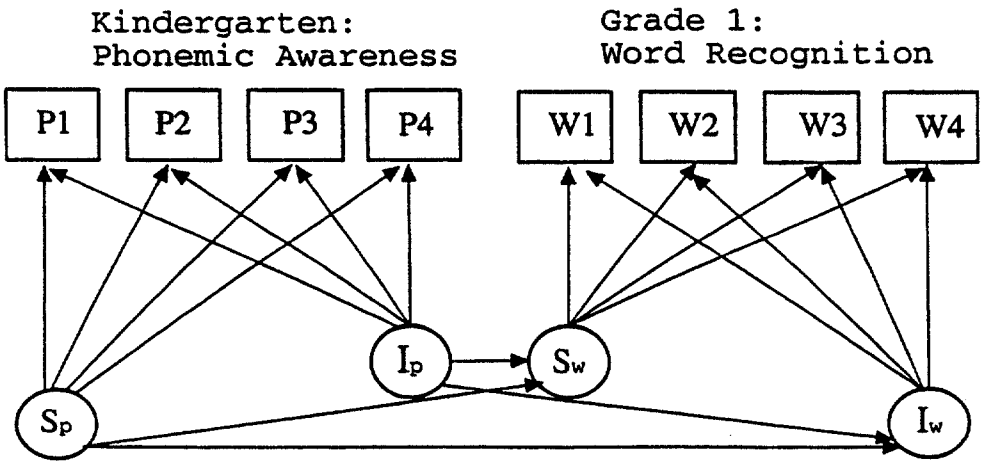
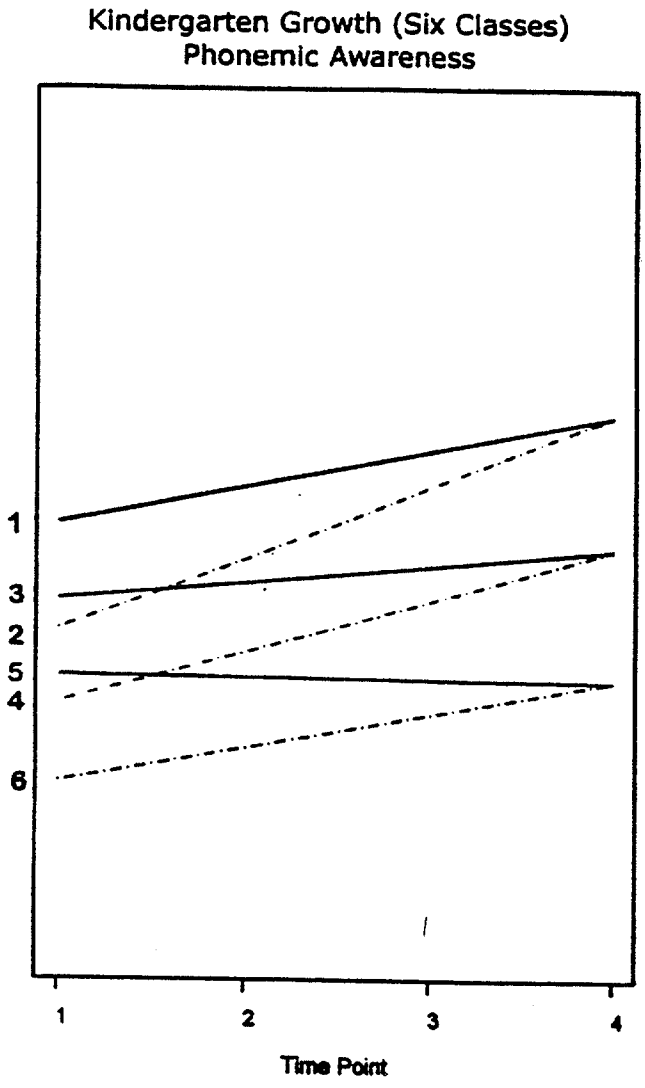


Figure 8.

Reading skills development: mixture model.


Kindergarten Growth (Six Classes)
Phonemic Awareness
1
3
2
5
4
6
1
2
3
4
Time Point

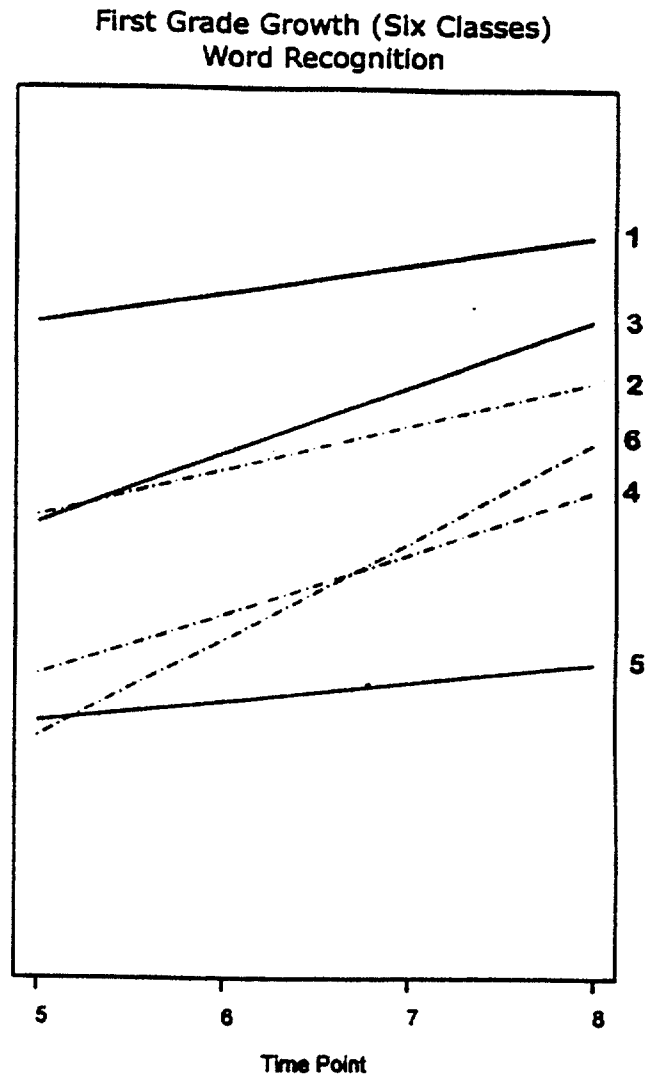
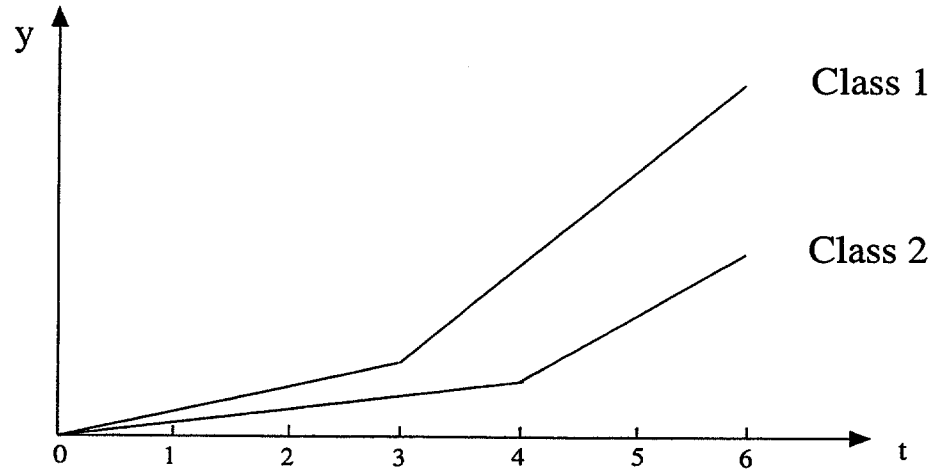
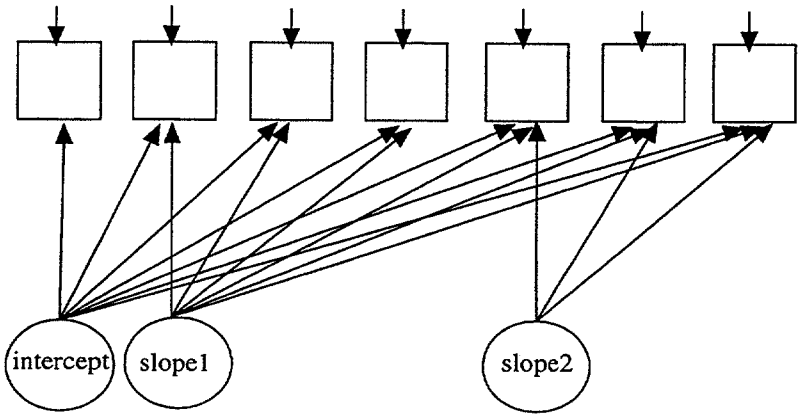

First Grade Growth (Six Classes)
Word Recognition
1
3
2
6
4
5
5
6
7
8
Time Point

Figure 9.

Piecewise growth modeling with individually-varying transition points.

Piecewise Growth Modeling: Individually-Varying Transition Points

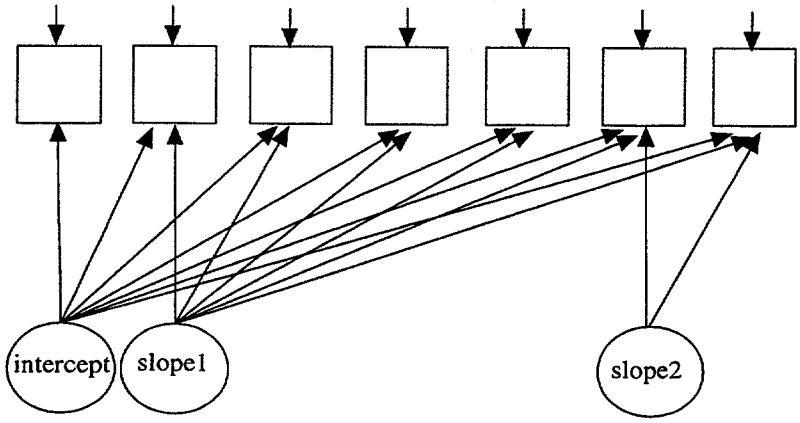




Class 1

slope1: 0,1,2,3,3,3,3

slope2: 0,0,0,0,1,2,3



Class 2

slope1: 0,1,2,3,4,4,4

slope2: 0,0,0,0,0,1,2

Figure 10.

Modeling of heavy drinking: predicting a distal outcome.

NLSY: Antecedents and Consequences

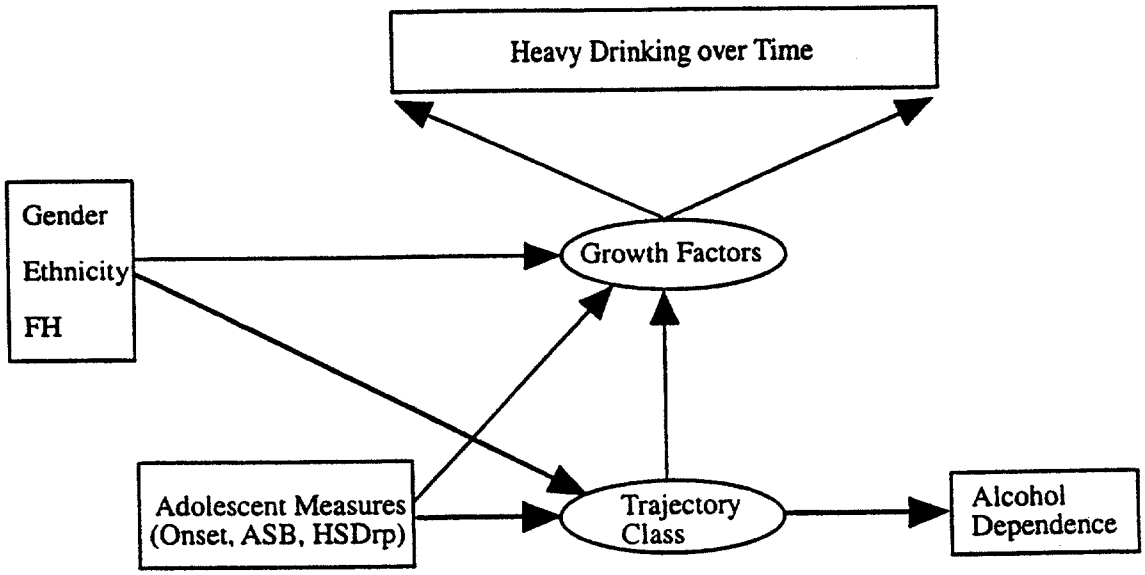
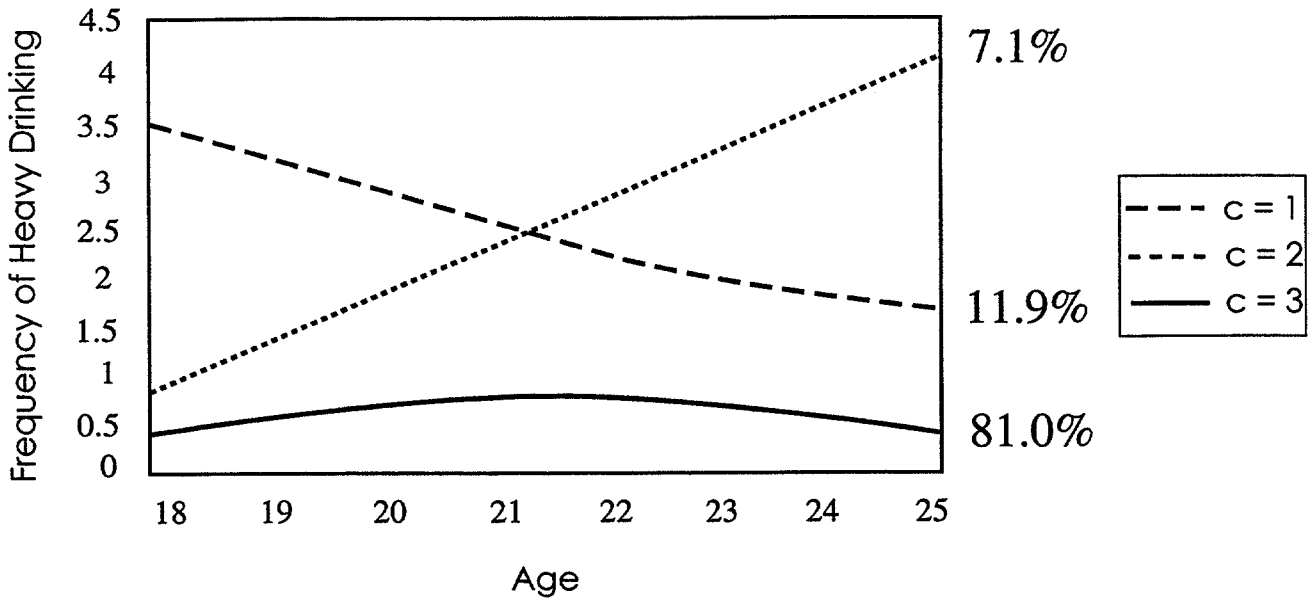


Figure 11.

Modeling of heavy drinking: prediction results.



	Odds Ratio
P(Dep I c=1) = 0.24	4.90
P(Dep I c=2) = 0.36	6.47
P(Dep I c=3) = 0.08	1.00