Psychological Methods
2003, Vol. 8, No. 3, 338–363


1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.3.338

# Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes

Daniel J. Bauer
North Carolina State University

Patrick J. Curran
University of North Carolina at Chapel Hill

Growth mixture models are often used to determine if subgroups exist within the population that follow qualitatively distinct developmental trajectories. However, statistical theory developed for finite normal mixture models suggests that latent trajectory classes can be estimated even in the absence of population heterogeneity if the distribution of the repeated measures is nonnormal. By drawing on this theory, this article demonstrates that multiple trajectory classes can be estimated and appear optimal for nonnormal data even when only 1 group exists in the population. Further, the within-class parameter estimates obtained from these models are largely uninterpretable. Significant predictive relationships may be obscured or spurious relationships identified. The implications of these results for applied research are highlighted, and future directions for quantitative developments are suggested.

Over the last decade, random coefficient growth modeling has become a centerpiece of longitudinal data analysis. These models have been adopted enthusiastically by applied psychological researchers in part because they provide a more dynamic analysis of repeated measures data than do many traditional techniques. However, these methods are not ideally suited for testing theories that posit the existence of qualitatively different developmental pathways, that is, theories in which distinct developmental pathways are thought to hold within subpopulations. One widely cited theory of this type is Moffitt's (1993) distinction between "life-course persistent" and "adolescent-limited" antisocial behavior trajectories. Moffitt's theory is prototypical of other developmental taxonomies that have been proposed in such diverse areas as developmental psychopathology (Schulenberg, O'Malley Bachman, Wadsworth, & Johnston, 1996; Zucker, 1986), social development (Brendgen, Vitaro, Bukowski, Doyle, & Markiewicz, 2001), and cognitive and language development (McCall, Appelbaum, & Hogarty, 1973; Rescorla, Mirak, & Singh, 2000). Statistical analyses conducted without attention to this heterogeneity may yield results that fail to accurately depict the relationships that hold within any one of the groups, including important predictive relationships (Jedidi, Jagpal, & DeSarbo, 1997; B. O. Muthén, 1989). There is thus a strong need for analytic methods that are capable of discerning and testing hypotheses about the developmental trajectories of unobserved population subgroups, or *latent trajectory classes.*

In response to this demand, promising new modeling techniques known as *growth mixture models* have recently been developed that permit investigators to estimate latent trajectory classes and to examine their unique relations to predictors or outcome measures (Li, Duncan, & Duncan, 2001; B. O. Muthén, 2001; B. Muthén & Shedden, 1999; Nagin, 1999; Nagin & Tremblay, 2001). The premise of these techniques is that the patterns in the repeated measures reflect a finite number of trajectory types, each of which corresponds to an unobserved or latent class in the population. The more general finite mixture model on

Daniel J. Bauer, Department of Psychology, North Carolina State University; Patrick J. Curran, Department of Psychology, University of North Carolina at Chapel Hill.

This work was funded in part by National Institute on Drug Abuse (NIDA) Fellowship DA06062 awarded to Daniel J. Bauer and Grant DA13148 awarded to Patrick J. Curran. We thank Kenneth Bollen, Andrea Hussong, Daniel Serrano, and the members of the Carolina Structural Equations Modeling Group for their valuable input throughout this project.

Correspondence concerning this article should be addressed to Daniel J. Bauer, Department of Psychology, College of Humanities and Social Sciences, North Carolina State University, Raleigh, North Carolina 27695-7801. E-mail: dan_bauer@ncsu.edu

which these techniques are based has a long history in the social sciences (e.g., Lazarsfeld, 1968; Quandt & Ramsey, 1978; Tukey, 1960). However, the underlying distributional assumptions of finite mixture models have not been thoroughly presented to applied psychological researchers, nor have the specific implications of these assumptions for growth mixture models been explored. We examine these issues more fully here.

One key point that we would like to highlight is that finite mixture models were developed for two purposes. The first corresponds to the motivations of most psychological researchers: to identify qualitatively distinct classes of individuals in the population of study. The second purpose, less well known to the applied researcher, is to approximate intractable or complex distributions with a small number of simpler component distributions. These two purposes of the model are quite distinct theoretically, but they are currently difficult to distinguish analytically. The implication of this point for growth mixture modeling is that the latent trajectory classes can often be interpreted in two very different ways. They may represent the trajectories of distinctive subgroups in the population of study, or they may provide an approximation to a complex but unitary population distribution of individual trajectories. In the latter case, it would be incorrect to interpret the classes as latent subgroups in the population. Doing so could lead to the identification of spurious effects or failure to identify relationships that are in fact significant. To make correct and valid inferences from data to theory, it is critically important to consider both possible interpretations of the latent trajectory classes when applying these models.

To clarify this issue, we first briefly review conventional growth modeling and its relation to growth mixture modeling. We then consider the analytical basis of the new techniques, the finite mixture model. Our short historical review of the development of finite mixture models reveals both the underlying assumptions of these models and two alternative perspectives on their application. A key point of this review is that the extracted components or classes may or may not reflect a heterogeneous population structure. We demonstrate this point with a simple example of a mixture of two univariate normal distributions. The dilemma for the applied researcher is that the fit statistics most commonly used to evaluate growth mixture models do not adequately discriminate between these two possibilities. We illustrate this point with a small simulation study. Our results suggest that despite the many new opportunities offered by these techniques, researchers should be cautious in the use and interpretation of growth mixture models, particularly when evaluating predictors of class membership. Strong substantive theory may provide important guidance on this matter, and new statistical developments may offer more formal means to distinguish these two conditions. We conclude by considering the implications of our results for applied research using these models and directions for future quantitative developments.

## Conventional Growth Models

It is useful to begin by considering the conventional random coefficient growth model because it represents a special case of a growth mixture model in which only one class is estimated. Conventional growth models can be estimated either as multilevel or hierarchical linear models or as structural equation models, and many excellent articles have been written on these techniques and their commonalities (e.g., see McArdle, 1988; McArdle & Epstein, 1987; Mehta & West, 2000; Meredith & Tisak, 1984, 1990; Raudenbush & Bryk, 2002, pp. 160–204; Willett & Sayer, 1994). We focus on the structural equation modeling approach, or the latent curve model.

At its core, the latent curve model is a confirmatory factor model with mean structure, where the factors represent the parameters of the individual trajectories. The mean and covariance structure of the repeated measures implied by the latent curve model are

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}\boldsymbol{\alpha} \tag{1}$$

and

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2}$$

where $\boldsymbol{\alpha}$ is a vector of latent variable means, or the means of the trajectory parameters. Individual variability in the trajectory parameters is captured in $\boldsymbol{\Psi}$, which holds the variances and covariances of the latent factors, and the residual variances and covariances of the repeated measures are contained in $\boldsymbol{\Theta}$. The elements of $\boldsymbol{\Lambda}$ define the shape of the latent trajectories and are often specified in advance to fit a particular functional form of growth.

A path diagram of a latent curve model specifying linear growth is given in Figure 1. Note that the manifest variables, $y_0$–$y_4$, represent the observed values of the repeated measures at five time points and are re-

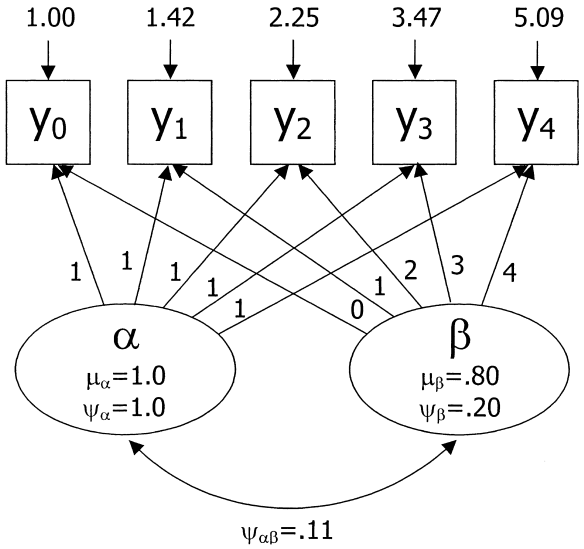


*Figure 1.* Path diagram of a single-group latent trajectory model. Displayed numbers are the population values of the parameters used in the simulation study.

lated to the latent intercept (α) and latent slope (β) through the parameterization of the factor loading matrix. To specify the linear functional form of growth, we fixed the factor loadings relating the repeated measures to the intercept factor at 1, and the factor loadings relating the repeated measures to the slope factor at values that increase linearly with time from 0 to 4 (here we make the unnecessary assumption of equally spaced time intervals). Typically, only the factor loadings are set to fixed values, and all other parameters in the model are estimated, with particular interest given to the estimated means, variances, and covariance of the growth factors.

It is often also of interest to predict individual variability in the trajectory parameters by one or more exogenous predictors. The model-implied means and covariance matrix for the repeated measures are then

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\kappa}) \tag{3}$$

and

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}' + \boldsymbol{\Psi})\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{4}$$

where $\boldsymbol{\kappa}$ is a vector of means for the predictors, $\boldsymbol{\Phi}$ is the covariance matrix of the predictors, and $\boldsymbol{\Gamma}$ is a matrix of regression coefficients capturing the effects of the exogenous variables on the latent trajectory parameters.

Several assumptions of the basic growth model are noteworthy. First, the trajectories of all individuals are assumed to be of the same functional form (e.g., linear). A second assumption is that the repeated measures data can be fully summarized by their means and covariances. To satisfy this condition, it is assumed that the repeated measures are multinormally distributed, implying that the individual growth parameters and time-specific residuals are also multinormal (Raudenbush, 2001, p. 514; Verbeke & Lesaffre, 1996; Verbeke & Molenberghs, 2000). Finally, when predictors are included, it is standard to assume that their effects are constant over the range of the trajectory parameter values (e.g., that the effects do not differ for individuals with low and high intercepts).

## Multiple-Groups Growth Modeling

In heterogeneous populations, the assumptions of the conventional growth curve model are generally untenable. Fortunately, they can be mitigated to some degree by using the multiple-groups framework in the structural equations approach to growth modeling (McArdle, 1989; B. O. Muthén & Curran, 1997). This approach differs in important ways from conventional growth models that incorporate the grouping variable as an exogenous predictor of the trajectory parameters. In the latter case, the groups differ only in mean levels of the growth parameters (B. O. Muthén, 1989). In contrast, in the multiple-groups approach, the grouping variable may be thought of as a moderator variable, and all of the model parameters potentially vary as a function of group membership. The basic model may be denoted by

$$\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_k\boldsymbol{\alpha}_k \tag{5}$$

and

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k, \tag{6}$$

where the subscript $k$ designates group and indicates that the parameters may differ over groups. A key focus of multiple-groups analysis is the test of model invariance over groups. For instance, one could test whether the groups differ only in mean levels of growth by constraining $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Psi}_k$, and $\boldsymbol{\Theta}_k$ to be equal over groups. The invariance of predictive relationships can also be tested in order to evaluate possible group by predictor interactions where the effect of the predictor differs significantly over groups (B. O. Muthén & Curran, 1997).

Although the multiple-groups framework is an important method for exploring population heterogene-

ity, its use is limited to the case in which the grouping variable is observed. However, it is often of interest to define taxonomic groups that are not known a priori on the basis of their developmental patterns. For instance, Moffitt's (1993) taxonomy of life-course persistent and adolescent-limited antisocial behavior trajectories is based on the presentation of antisocial behavior across the life span. In this case, group membership is unobserved and must instead be inferred from the repeated measures themselves.

## Growth Mixture Models

Growth mixture models generalize the multiple-groups framework to the case in which the grouping variable is either completely unobserved or missing for some portion of the cases (B. O. Muthén, 2001; B. Muthén & Shedden, 1999; Nagin, 1999; Nagin & Tremblay, 2001; Verbeke & Lesaffre, 1996). In this case, the grouping variable is replaced by a probability of class membership, and each case contributes to the parameter estimates of each latent class commensurate with its probability of membership in that class (L. K. Muthén & Muthén, 1998, Appendix 8). Because group membership is unobserved, the proportion of cases in each latent class is unknown and must be estimated along with the other parameters of the model.

The uncertainty of group membership introduces a number of complexities into the estimation of the model. It is thus common to impose further restrictions on the structure of the model. For instance, the functional form of growth is often held to be invariant over groups (i.e., $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}$). The variance components of the model may also be held invariant over groups (i.e., $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}$), again implying that the classes differ only in their mean trajectories (B. O. Muthén, 2001; Verbeke & Lesaffre, 1996). Another common approach involves constraining the variance/covariance matrix of the growth factors to zero (i.e., $\boldsymbol{\Psi}_k = \mathbf{0}$; Nagin, 1999; White, Johnson, & Buyske, 2000). This last model implies that all of the individual variability in growth is captured by the class mean trajectories (i.e., fixed effects) and that any individual deviations from the class mean trajectories are random error. Often these constraints are imposed for statistical expediency rather than from substantive theory, and in practice they are often rejected when tested, suggesting potential model misspecifications. We thus center our discussion on the more general model with nonzero variance components that vary over classes.

Predictors can be incorporated into this model in two different ways. First, exogenous variables may be used to predict within-class variability in the latent trajectory parameters. This parallels the evaluation of predictors in multiple-groups models and permits one to test whether the effects of the predictors vary over classes.[1] Given the uncertainty of group membership, it may also be of theoretical interest to predict the probability that an individual belongs to a particular group. This constitutes the second method for modeling predictors and is accomplished through the inclusion of a multinomial regression submodel that relates the predictors to the individual probabilities of group membership (Jones, Nagin, & Roeder, 2001; B. Muthén & Shedden, 1999).

A key complication that arises with these models is that the number of latent trajectory classes must be specified, not estimated. Given that the number of classes is seldom known with certainty in advance, it is common for investigators to fit several models with different numbers of classes. Model fit statistics such as Akaike's information criterion (AIC), the Bayesian information criterion (BIC) or the consistent AIC (CAIC) may then offer a guide for selecting the model with the optimal number of classes (see Bozdogan, 1987; McLachlan & Peel, 2000, pp. 203, 207–208; L. K. Muthén & Muthén, 1998, pp. 371–372; Nagin, 1999). These fit indices are based on the value of the likelihood function, so they reward models that more accurately reproduce the observed data, but they also exact a penalty for the number of parameters in the model, favoring models with fewer trajectory classes. In a review of the performance of these statistics, McLachlan and Peel (2000) noted that the AIC tends to overestimate the number of classes present, whereas the BIC (and by extension the CAIC) may underestimate the number of classes present, particularly in small samples.

Alternative fit statistics have been proposed for selecting the true number of classes; these alternative fit statistics reward models that produce well-separated clusters, that is, models in which the estimated prob-

[1] Note that this model resembles the switching regression model of Quandt and Ramsey (1978) in which latent classes are defined on the basis of their unique relations to predictor variables. Similarly, in conditional growth mixture models, classes may be defined not only by the patterns present in the repeated measures but also by distinctive within-class relations to predictor variables.

abilities of group membership approach one or zero (see McLachlan & Peel, 2000, for a review). These statistics include the normalized entropy criterion (NEC; Biernacki & Govaert, 1999; Celeux & Soromenho, 1996) and the classification likelihood criterion (CLC; Biernacki & Govaert, 1997). Finally, the integrated completed likelihood criterion (ICL and ICL–BIC; Biernacki, Celeux, & Govaert, 2000) penalizes for both the number of parameters and the quality of classification, making it a more conservative selection criterion (see the Appendix for further details on these fit measures).

## Finite Mixture Models

The analytical basis of the growth mixture model is the finite mixture model. Finite mixture models have a long history (e.g., Pearson, 1894), and they have been popularized in the social sciences by specific models such as switching regression (Quandt & Ramsey, 1978) and latent class analysis for binary indicators (Clogg, 1995). More recently, finite mixture models have been extended to confirmatory factor analysis (Blåfield, 1980; Yung, 1997) and structural equation models (Arminger, Stein, & Wittenberg, 1999; Jedidi et al., 1997), making them sufficiently general to be useful in a variety of applied psychological research. Just as important, software is now readily available for estimating these models, including Mplus (L. K. Muthén & Muthén, 1998), the MECOSA program for GAUSS (Arminger, Wittenberg, & Schepers, 1996), and Mx (Neale, Boker, Xie, & Maes, 1999), with varying degrees of flexibility. With such software it is possible to estimate mixtures of normal distributions with structured means and covariances, including growth mixture models.[2] Yet despite the wide availability of this software, to our knowledge the basic assumptions of finite normal mixture models have not been explicated for the applied researcher. We begin our examination of these assumptions by considering a two-class normal mixture for a single variable. This model is simple in comparison to the growth mixture model, but it provides a suitable basis for understanding the analytical properties of finite normal mixture models more generally.

An example of a two-class univariate normal mixture model is displayed in Figure 2. The probability density function for this model is

$$f(x) = p_1 g_1(x) + (1 - p_1) g_2(x), \tag{7}$$

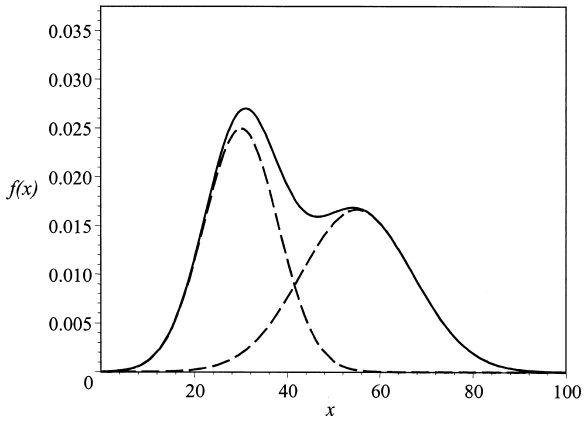


*Figure 2.* The probability distribution of a two-component mixture of normal distributions. The solid line depicts the distribution of the aggregate mixture; dashed lines indicate the distributions of the components.

where $f(x)$ is the aggregate or composite density function, $g_1(x)$ and $g_2(x)$ are the component densities of the two classes, and $p_1$ and $(1 - p_1)$ are the mixing proportions that designate the number of cases originating from each component. The component densities are defined to be normally distributed so that

$$g_1(x) = (2\pi\sigma_1^2)^{-1/2} e^{-[(x_i-\mu_1)^2/2\sigma_1^2]}$$

$$g_2(x) = (2\pi\sigma_2^2)^{-1/2} e^{-[(x_i-\mu_2)^2/2\sigma_2^2]}. \tag{8}$$

Note that the density function for each component in Equation 8 is completely described by two parameters, the class mean and class variance. To fit the model in Equation 7, it is thus necessary to estimate five parameters, $\mu_1$, $\sigma_1^2$, $\mu_2$, $\sigma_2^2$, and $p_1$. Each parameter is subscripted by $k = 1,2$ to designate the component the parameter references.

What is not always appreciated about this model is that nonnormality of $f(x)$ is a necessary condition for estimating the parameters of the normal components $g_1(x)$ and $g_2(x)$. Indeed, it was recognition of this fact

[2] Several other software modules are currently capable of estimating growth mixture models under the specific constraint that there be no individual variability in growth within classes (i.e., $\boldsymbol{\Psi}_k = \mathbf{0}$). These include the PROC TRAJ macro for SAS (Jones et al., 2001), the MMLCR library for Splus (White et al., 2000), and Latent GOLD (Vermunt & van Dijk, 2001). Though these programs do not currently allow for the estimation of individual variability in growth within classes, they do permit the disturbances to follow alternative distributions to the normal.

that led Pearson (1894) to develop an estimation approach for normal mixture models using the method of moments. Pearson (1894) began by showing that if the class proportions, means, and variances were known, they could be used to solve for the higher order moments of the aggregate data. For instance, the mean, variance, and third and fourth central moments (relating to skew and kurtosis, respectively) of $f(x)$ can be expressed as

$$\mu = p_1\mu_1 + (1 - p_1)\mu_2, \tag{9}$$

$$\sigma^2 = p_1(\sigma_1^2 + d_1^2) + (1 - p_1)(\sigma_2^2 + d_2^2), \tag{10}$$

$$m^3 = p_1 d_1(3\sigma_1^2 + d_1^2) + (1 - p_1)d_2(3\sigma_2^2 + d_2^2), \tag{11}$$

and

$$m^4 = p_1(3\sigma_1^4 + 6\sigma_1^2 d_1^2 + d_1^4) + (1 - p_1)(3\sigma_2^4 + 6\sigma_2^2 d_2^2 + d_2^4), \tag{12}$$

where $d_1 = \mu_1 - \mu$ and $d_2 = \mu_2 - \mu$. The task of estimation then involves reversing the unknowns in these equations. The higher order moments of the observed data (up to the fifth central moment) are substituted for their implied values, and the system of equations is solved to obtain estimates for $p_1$, $\mu_1$, $\mu_2$, $\sigma_1^2$, and $\sigma_2^2$ (see A. C. Cohen, 1967, for further detail).

The same equations indicate that a mixture of normals is, except in certain "degenerate" cases, necessarily nonnormal. The first degenerate case is if $p_1 = 1$. The second component is then superfluous, and Equation 7 can be reduced to $f(x) = g_1(x)$, or essentially a one-component solution. In this case the parameters of the second component cannot be identified; because no cases arise from the second component, there is no information with which to estimate its mean or variance. The second degenerate case is where the two components completely overlap, or where $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$. In this case, we could substitute $g_1(x)$ for $g_2(x)$ in Equation 7 to again arrive at the essentially one-class model $f(x) = g_1(x)$. In this case, we cannot identify the relative proportions of cases in each component, because the two components are indistinguishable. What these degenerate cases demonstrate is that if the aggregate distribution is normal, then only one normal component is necessary to summarize the data, and there is no remaining information with which to identify a second normal component. A more formal proof of this point has been given by Teicher (1960, pp. 63–68). The same argument also holds at the multivariate level: If the aggregate distribution is multivariate normal, it may be fully characterized by a single mean vector and covariance matrix, and no more than one component can be identified. Nonnormality of the aggregate distribution is thus critical for obtaining a nontrivial solution for the mixture.

Not only is nonnormality *required* for the solution of the model to be nontrivial, it may well also be a *sufficient* condition for extracting multiple components. For this reason, mixtures of normal distributions are often used to provide an approximation to complex or intractable distributions (Escobar & West, 1995; Ferguson, 1983; Roeder & Wasserman, 1997; Sorenson & Alspach, 1971; for a review, see Everitt & Hand, 1981, pp. 118–124; McLachlan & Peel, 2000, pp. 7–8; Titterington, Smith, & Makov, 1985, pp. 18–34). Presaging this use of normal mixture models, Pearson (1894, p. 72) wrote, "Even where the material is really homogeneous, but gives an abnormal frequency-curve, the amount and direction of abnormality will be indicated if this frequency-curve can be split up into normal curves." When used to this end, the characteristics of the component distributions are of little intrinsic interest, because they serve only as analytical devices for examining the aggregate distribution. This is in direct contrast to the use of mixtures for identifying population heterogeneity, where the characteristics of the component distributions are of paramount interest and the aggregate distribution is typically of concern only insofar as it may be used to identify those characteristics.

It is also possible to view these two uses of normal mixtures as alternative explanations for the results of a given model: Do the components represent true latent subgroups in the population, or are they serving only to approximate what is in fact a homogeneous but nonnormal distribution? This is a question that Pearson (1895) himself posed over a century ago:

> The question may be raised, how are we to discriminate between a true curve of skew type and a compound curve [or mixture], supposing we have no reason to suspect our statistics *a priori* of mixture. I have at present been unable to find any general condition among the moments, which would be impossible for a skew curve and possible for a compound, and so indicate compoundness. I do not, however, despair of one being found. (p. 394)

Despite Pearson's optimism, the ensuing century of research on these models has not yet provided an unequivocal answer to this basic question. The dilemma for the applied researcher is then to determine which of these two explanations is most reasonable

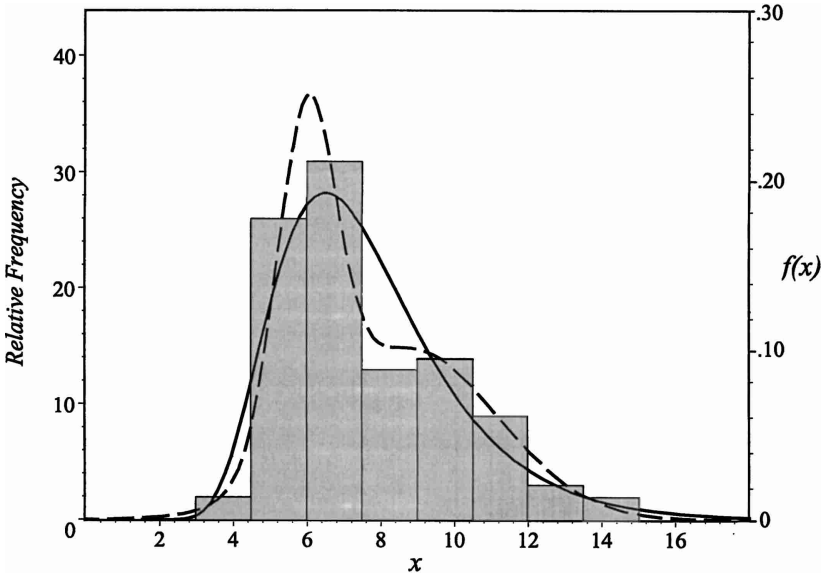


*Figure 3.* The sample histogram displays the relative frequency of values of variable *x* for 100 cases. Overlaid on the histogram are the probability distributions for a two-component normal mixture (dashed line) and a lognormal distribution (solid line). Both functions appear to fit the data well, but the population generating function was lognormal.

for making inferences about the structure of the population.

To make this dilemma clear, let us consider the sample data displayed in Figure 3 (after Titterington et al., 1985, p. 30). The histogram shows evidence of positive skew, and it appears to have two modes. Two probability density functions are superimposed on the distribution. The dashed line is the density function of a two-component normal mixture like the one described in Equation 7. From this model, the researcher might be tempted to conclude that the sample data arise from two unobserved groups, one large group with a mean of around 6 and another smaller group with a mean of around 10. The alternative model represented by the solid line is that the distribution is lognormal and from a single group. The two models appear to fit the observed data equally well (though in fact it is the single lognormal distribution that is correct). Without strong theoretical justification for choosing one distribution over another, it may be quite difficult to determine empirically which model reflects the true population structure.[3] Far from being a trivial matter, model selection may have a substantial impact on further analyses. These concerns apply equally to multivariate normal mixtures with structured means and covariances, including growth mixture models, as we now demonstrate.

## Distributions and Latent Trajectory Classes

Our examination of the univariate normal mixture indicated that nonnormality of the aggregate distribution is critical for the estimation of the latent component distributions. By extension, we should expect multivariate nonnormality to play a key role in the estimation of latent trajectory classes in growth mixture models. Of course, as Pearson (1894) noted, this nonnormality may indeed reflect the mixing of several unobserved groups, in which case it would be of great importance to use these techniques to explicitly model this heterogeneity. However, it is also possible that the individual variation follows an alternative distribution function, or that the nonnormality is a consequence of poor measurement scaling (among other possibilities), and that only one group actually exists

[3] As a case in point, it is interesting that Pearson (1894) originally demonstrated his method for estimating a univariate normal mixture with data collected by W. F. Raphael Weldon on the carapace lengths of crabs. Weldon was skeptical of the validity of Pearson's two-component model, pointing out several alternative mechanisms that, in his view, might better account for the nonnormal shape of the empirical distribution (Stigler, 1986, pp. 336–338).

in the population. Given that the primary use of these models in psychology and the social sciences has been to identify population heterogeneity, it is critical to know whether a multiple-class model could be estimated from the nonnormal data of a single group, whether it would fit the data better than a single-class model, and what the implications would be of selecting it over the single-class model. We consider these issues from the perspective of an investigator seeking to model population heterogeneity, whose primary interests are to identify and make valid inferences about possible subgroups in the population and whose aim is not simply to find an approximation for a nonnormal but homogeneous distribution of repeated measures.

Drawing on the statistical theory discussed above, we generated three key hypotheses. Our first hypothesis centered on the role of nonnormality in the estimation of multiple trajectory classes. For the simpler univariate normal mixture, we noted that nonnormality is a necessary condition for the extraction of multiple latent components or classes. By extension, if the distribution of the repeated measures is multivariate normal, it should not be possible to obtain a nondegenerate multiclass solution when fitting a correctly specified growth model. That is, with the correct model (e.g., functional form), the implied means and covariances of a single class should fully reproduce the observed distribution, and additional classes should not be necessary (nor would there be enough information with which to identify their parameters). In finite samples, however, random variability ensures that no distributions are truly normal, and thus it may be possible to obtain a nondegenerate multiclass solution even with data drawn from a multivariate normal population distribution. Because sampling variability is greatest in small samples, we expected that this would occur more often for small than for large samples. Conversely, we expected that with nonnormally distributed repeated measures, proper multiclass solutions would be the rule even when only one class actually existed. Further, in this case, increased sample sizes should function primarily to increase the information available for analysis and thus facilitate convergence to a proper solution.

We next considered the conditions under which a multiclass model would fit the data better than the correct single-class model. If the data are generated from a single multivariate *normal* distribution, model fit statistics that reward parsimony, such as the AIC, the BIC, and the CAIC, should reliably indicate that only one class is necessary to reproduce the data, because any additional classes serve only to capture sampling variability. Further, we expected that the components estimated from this data would overlap considerably, so that the NEC, the CLC, and the ICL–BIC would also reject multiple-class models (see the Appendix for details on these fit measures). We expected the opposite pattern of results for data drawn from multivariate *nonnormal* distributions. Recalling that one function of normal mixtures is to approximate complex or unknown distributions, we hypothesized that a multiple-class model would perform substantially better than a single-class model at reproducing nonnormally distributed repeated measures and that this would be reflected in model fit statistics such as the AIC, the BIC, and the CAIC. In addition, we expected the degree of separation between the classes to increase with the degree of nonnormality of the data, so that the NEC, the CLC, and the ICL–BIC would also favor multiple-class models. Our second hypothesis was thus that multiclass models would optimally fit repeated measures drawn from a multivariate nonnormal distribution even if only one group actually existed in the population.

Our third hypothesis concerned the implications of fitting a single-class versus a multiclass model to nonnormal data when only one group actually existed in the population. In both cases the model is misspecified. Fitting the single-class model involves a violation of the assumption of multivariate normality. However, the maximum likelihood function used to fit these models is known to produce consistent parameter estimates even when data are nonnormal, and robust standard errors may be calculated (Bollen, 1989, pp. 416–418; Browne, 1984; Satorra & Bentler, 1994). Alternatively, as the second hypothesis indicated, we expected that multiclass models would perform much better at capturing the nonnormality of the repeated measures. However, if a mixture is used solely for approximation purposes, the parameters of the components are usually of little intrinsic interest because they have no analog in the population. That is, we are estimating parameters that do not really exist in the population and are hence largely uninterpretable. What then would be the implication of mistakenly concluding that the estimated classes represent latent subgroups and proceeding to interpret these parameters? We expected that by centering our attention on the within-class estimates, the true role of exogenous predictors of individual change could go undetected or spurious relationships could be identified.

We empirically evaluated these hypotheses with a small simulation study designed to reflect conditions that might commonly be encountered by investigators seeking to use these models in practice (e.g., moderately large samples, mildly nonnormal data, a moderate number of assessment occasions). We stress that we intend this study not to be a comprehensive evaluation of all such conditions but to provide a clear empirical demonstration of the validity of the proposed hypotheses.

### *Simulation Design*

All data were generated to be consistent with the five-occasion linear growth model shown in Figure 1. It is important to note that in all conditions, only a single homogeneous group existed in the population. The population mean trajectory was parameterized so that, on average, scores would increase over time ($\mu_\alpha = 1.00$ and $\mu_\beta = 0.80$). The variance components were specified to provide meaningful individual variability in intercepts and slopes ($\psi_\alpha = 1.00$ and $\psi_\beta = 0.20$). Further, intercepts and slopes were positively correlated to a modest degree ($\psi_{\alpha\beta} = 0.11$; $\rho_{\alpha\beta} = 0.25$). The total variance of each of the five repeated measures was partitioned equally among the underlying trajectory and the time-specific residuals (e.g., $\rho_t^2 = .50$).

Five hundred samples at each of two sample sizes, $N = 200$ and $N = 600$, were generated for three distributional conditions. In the first condition, the data were generated to be normally distributed (i.e., with univariate skew 0 and kurtosis 0). The other two conditions involved transformations of the repeated measures data using Fleishman's (1978) method for generating nonnormal random variables, as extended by Vale and Maurelli (1983) and implemented in EQS (Bentler, 1995). Specifically, in these two conditions, the repeated measures data were transformed to have univariate skew 1 and kurtosis 1, and skew 1.5 and kurtosis 6, respectively. These values are well within the range of skew and kurtosis encountered in applied psychological research (Micceri, 1989), and they represent minor deviations from normality that would typically be of little concern for conventional (one-class) growth modeling. Histograms displaying the shapes of these distributions are displayed in Figure 4.

To test our hypotheses, we estimated one- and two-class models for the data. We used Mplus 2.01 to estimate the models, employing the EM estimator with the MLR option to obtain robust standard errors (L. K. Muthén & Muthén, 1998). A modified version of the RUNALL utility was used to compile the results (Nguyen, Muthén, & Muthén, 2001). We did not examine models with more than two classes, because our hypotheses concerned only whether more than one class could be extracted and provide better fit to the data. Finite normal mixture models are known to have poorly behaved likelihood functions, potentially including many local solutions, and, when the variance parameters are permitted to vary over classes, singularities at the edges of the parameter space that can lead to nonconvergence (McLachlan & Peel, 2000, pp. 94–97). For this reason, two-class models were estimated both with and without across-class equality constraints on the variance components (e.g., $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}$).[4] Next, given the possibility that there would be multiple local solutions, all two-class models were estimated with six separate sets of start values. One set of start values was based on the recommendation that start values for multiclass models should be derived from the parameter estimates obtained from single-group models (L. K. Muthén & Muthén, 1998, p. 132). Following this recommendation, we used the single-group population parameter values as start values for all of the parameters except the growth factor means, which were set higher in one group than the other for both growth factors ($\hat{\mu}_\alpha = 1.50$ and $\hat{\mu}_\beta = 1.60$ for Class 1, and $\hat{\mu}_\alpha = 0.00$ and $\hat{\mu}_\beta = 0.00$ for Class 2). The other five sets of start values were generated randomly.[5] Our use of random start values is consistent with their use in other simulation studies on finite normal mixtures (e.g., Biernacki, Celeux, & Govaert, 1999; McLachlan & Peel, 2000, p. 217).

[4] Although they are statistically expedient, we do not regard these equality constraints as optimal from a theoretical standpoint, and in our experience, they are rarely found to be tenable in practice. Indeed, implementing these constraints is in some ways inconsistent with the spirit of the analysis, because one is forcing the majority of the parameter estimates to be the same over classes (permitting only mean differences in the within-class trajectories). Further, McLachlan and Peel (2000, pp. 97–98) have cautioned that "imposition of the constraint of equal component-covariance matrices can have a marked effect on the resulting estimates and the implied clustering."

[5] The start value for each parameter was obtained by taking a random draw from a normal distribution with mean equal to the single-group population value for the parameter and a standard deviation set to provide broad coverage of the surrounding parameter space.

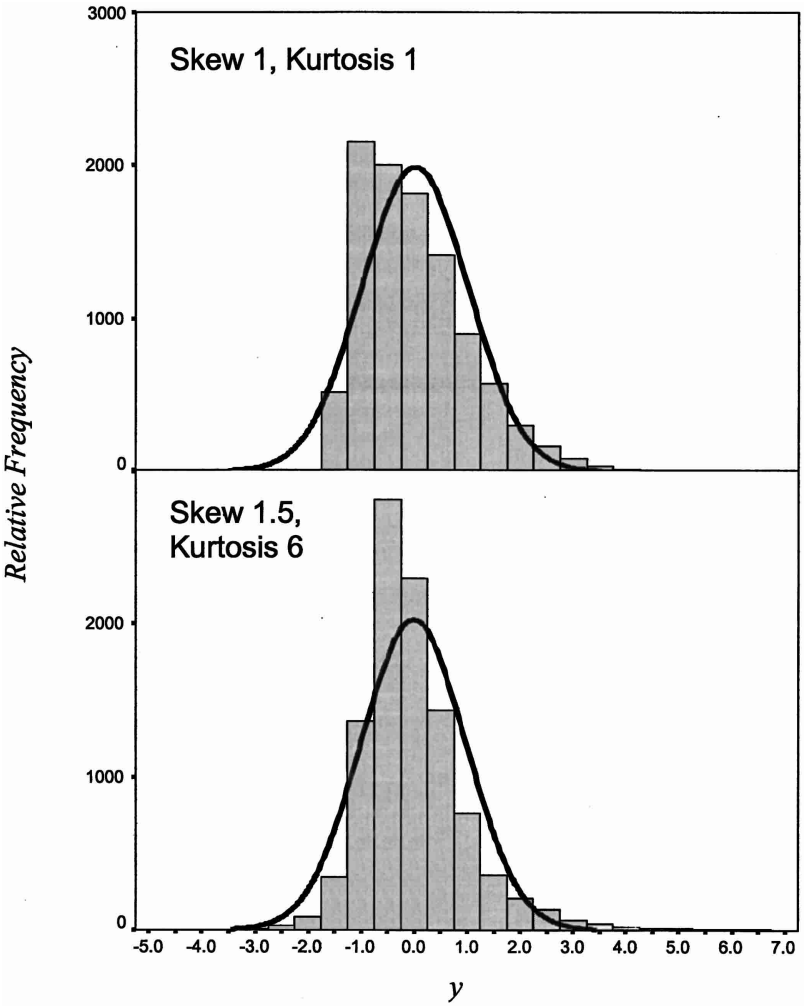


*Figure 4.* Example histograms displaying the shape of the three distributions that were used in the simulation study for standardized data at $N = 10{,}000$. For comparison, the probability function of a normal distribution is overlaid on the histograms.

The model was allowed 1,000 iterations to converge. We adopted the following algorithm for selecting solutions for analysis:

1. When a given replication failed to converge with any of the six sets of start values, the solution was labeled nonconvergent.

2. When more than one set of start values led to convergence for a given replication, the solution with the maximum (best) log-likelihood was selected. This again follows standard practice in studies of finite normal mixtures (Biernacki et al., 1999; Everitt & Hand, 1981; McLachlan & Peel, 2000, p. 217).

3. The solution selected from Step 2 was considered improper if any of the parameter estimates fell outside of their permissible boundaries (i.e., negative variances, or correlations greater than one).

Unless convergence was of explicit interest, nonconverged and improper solutions were excluded from the analyses because such solutions are rarely interpreted in practice (Chen, Bollen, Paxton, Curran, & Kirby, 2001). However, additional analyses were also conducted that included improper solutions, and the results did not deviate in any meaningful way from those reported here.

Table 1
*Convergence Rate of Two-Class Models (of 500 Samples) by Model Parameterization, Sample Size, and Degree of Nonnormality*

| Sample size and distribution | Failed to converge | Converged | |
|---|---|---|---|
| | | Improper solution | Proper solution |
| Class-invariant variance and covariance parameters | | | |
| $N = 200$ | | | |
| Skew 0, kurtosis 0 | 6 (1) | 193 (39) | 301 (60) |
| Skew 1, kurtosis 1 | 0 | 232 (46) | 268 (54) |
| Skew 1.5, kurtosis 6 | 0 | 150 (30) | 350 (70) |
| $N = 600$ | | | |
| Skew 0, kurtosis 0 | 23 (5) | 108 (22) | 369 (74) |
| Skew 1, kurtosis 1 | 0 | 211 (42) | 289 (58) |
| Skew 1.5, kurtosis 6 | 0 | 74 (15) | 426 (85) |
| Class-varying variance and covariance parameters | | | |
| $N = 200$ | | | |
| Skew 0, kurtosis 0 | 4 (1) | 450 (90) | 46 (9) |
| Skew 1, kurtosis 1 | 0 | 170 (34) | 330 (66) |
| Skew 1.5, kurtosis 6 | 0 | 162 (32) | 338 (68) |
| $N = 600$ | | | |
| Skew 0, kurtosis 0 | 31 (6) | 380 (76) | 89 (18) |
| Skew 1, kurtosis 1 | 0 | 29 (6) | 471 (94) |
| Skew 1.5, kurtosis 6 | 0 | 15 (3) | 485 (97) |

*Note.* Values in parentheses are percentages.

### *Nonnormality and the Estimation and Fit of Latent Trajectory Classes*

One important question was whether or not a two-class latent trajectory model could be estimated from data generated from a population consisting of a single group. On the basis of the theoretical model, we expected that two-class models fit to data drawn from a multivariate normal distribution would perform poorly, often failing to converge or settling on an improper solution. Conversely, we expected that the convergence rate for two-class solutions would significantly increase as the degree of nonnormality in the observed data increased.

Both hypotheses were empirically supported. As Table 1 indicates, two-class models failed to converge more often when fit to data drawn from a normal as opposed to a nonnormal population distribution regardless of whether equality constraints were placed on the variance and covariance parameters over classes.[6] Further, as we had anticipated, the rate of nonconvergence for both models was lower for the $N = 200$ condition (occurring in about 1% of samples) than for the $N = 600$ condition (occurring in 5%–6% of samples), an effect that may be understood as a consequence of greater sampling variability from the population distribution. As the sample size increases (e.g., at $N = 600$) the convergence rate declines because the empirical distribution begins to obtain its asymptotically multivariate normal form. In contrast, the rate of convergence was 100% in the nonnormal conditions regardless of sample size.

Comparing the models with and without equality

[6] The rate of nonconvergence in the normal condition was much higher for any single set of start values. When equality constraints were placed on the variance and covariance parameters, the rate of nonconvergence for each of the six sets of start values ranged from 5% to 15% of replications at $N = 200$ and from 24% to 33% of replications at $N = 600$. When these constraints were relaxed, similar rates of nonconvergence were obtained, ranging from 8% to 11% at $N = 200$ and from 24% to 31% at $N = 600$. In contrast, in the nonnormal conditions, most sets of start values lead to near 100% convergence rates, particularly when no equality constraints were placed on the variance or covariance parameters. This again illustrates the greater difficulty of obtaining convergence in the normal condition, because it was in this condition especially that multiple sets of start values were necessary to find a solution.

constraints reveals three primary findings. First, releasing the equality constraints had little effect on the proportion of models failing to converge, suggesting that in this case these constraints were largely unnecessary to avoid singularities on the likelihood surface (their primary purpose). Second, the constraints had a large impact on the proportion of proper to improper solutions that were obtained. Imposing equality constraints on the variance and covariance parameters greatly facilitated obtaining a proper solution in the normal condition. For instance, in the normal condition at $N = 600$, proper solutions were obtained in 74% of samples when the equality constraints were imposed but in only 18% of samples when they were not. This low rate of proper solutions in the unconstrained model may reflect the fact that there is simply not much information with which to estimate the additional variance and covariance parameters of the model. The opposite results were obtained for the nonnormal conditions. In the skew 1, kurtosis 1 condition at $N = 600$, 58% of samples yielded proper solutions with the constrained model, compared with 94% of samples when these constraints were relaxed. The skew 1.5, kurtosis 6 condition showed a similar, though less pronounced, pattern. Finally, likelihood ratio tests between the constrained and unconstrained models supported the retention of equality constraints in the majority of samples in the normal condition (82% of the time at $N = 200$; 66% at $N = 600$) but rejected them almost 100% of the time in the nonnormal conditions.[7] Given this finding, we considered both the constrained and unconstrained models for the normal condition but restricted our analysis to the unconstrained model for the nonnormal conditions.

Given that two-class models can be successfully fit to data drawn from a homogeneous population distribution, our second question concerned the conditions under which the two-class model would be identified as optimal. For replications with proper solutions, we examined the comparative fit of the one- and two-class models using the AIC, BIC, CAIC, NEC, CLC, and ICL–BIC. We had hypothesized that these fit statistics would correctly point to the presence of a single group only in the case where the repeated measures were generated from a multivariate normal distribution. Supporting this prediction, in the normal condition, the average values of these statistics were consistently higher for the two-class model than for the one-class model, signifying the superiority of the one-class model. This was true at both sample sizes and regardless of whether equality constraints were placed on the variance and covariance parameters of the two-class model, as can be seen in Tables 2 and 3. Of the fit statistics, the AIC showed the worst performance, indicating that the two-class model was superior in 21%–33% of replications. This finding is consistent with the known tendency for the AIC to indicate that too many classes should be estimated (Celeux & Soromenho, 1996; Soromenho, 1993). The other fit statistics performed quite well in rejecting the two-class model. Thus, if the population truly consists of a single homogeneous group, and the data are sampled from a multivariate normal distribution, then it is unlikely that a two-class model will fit the data better than the correct population model. These results are encouraging, because they indicate that under these conditions, the fit statistics (with the exception of the AIC) are leading to correct inferences about the number of latent subgroups in the population.

In contrast, in the conditions in which the data were drawn from nonnormal distributions, the two-class model usually converged on a proper solution that fit the data better than the single-class model. In this case, the AIC, the BIC, and the CAIC supported selection of two classes in almost 100% of the replications at both sample sizes. As hypothesized, the amount of improvement in fit increased with the degree of nonnormality of the repeated measures. For example, in the skew 1, kurtosis 1 condition, moving from one to two classes resulted in an average improvement of 2.4% in the BIC at $N = 200$, increasing to 2.8% at $N = 600$. Likewise, in the skew 1.5, kurtosis 6 condition, moving from one to two classes improved the BIC by 3.8% at $N = 200$ and by 4.4% at $N = 600$. In empirical applications, comparably sized improvements in the BIC have been interpreted as evidence of multiple trajectory classes (e.g., Hill, White, Chung, Hawkins, & Catalano, 2000; Li et al.,

[7] A large part of the reason for this difference is that in the normal case, even when two classes were estimable, their parameter estimates were often quite similar (approaching the degenerate case in which a normal aggregate distribution is resolved into two identical components with arbitrary proportions). Hence, constraining some of these estimates to be the same over classes has little impact on the fit of the model. In contrast, as shown in Table 5, the parameter estimates for the two classes in the nonnormal conditions were quite distinctive, so constraining them resulted in large and significant decrements in model fit.

Table 2
*Relative Fit of One-Class Versus Two-Class Models for Proper Solutions (of 500 Samples, N = 200)*

| Fit statistic | % of time statistic favors two-class model | Mean difference[a] | Mean % change in fit statistic[a] |
|---|---|---|---|
| Skew 0, kurtosis 0: Class-invariant variance and covariance parameters (301 of 500 samples) | | | |
| AIC | 25.58 | −1.32 | −0.03 |
| BIC | 0.66 | −11.21 | −0.27 |
| CAIC | 0.33 | −14.21 | −0.35 |
| NEC | 5.98 | −36.65 | −3,665.09 |
| CLC | 5.98 | −97.93 | −2.42 |
| ICL–BIC | 0 | −113.82 | −2.77 |
| Skew 0, kurtosis 0: Class-varying variance and covariance parameters (46 of 500 samples) | | | |
| AIC | 32.61 | −2.04 | −0.05 |
| BIC | 0 | −38.33 | −0.94 |
| CAIC | 0 | −49.33 | −1.20 |
| NEC | 0 | −6.93 | −692.75 |
| CLC | 0 | −118.23 | −2.93 |
| ICL–BIC | 0 | −176.51 | −4.31 |
| Skew 1, kurtosis 1: Class-varying variance and covariance parameters (329 of 500 samples)[b] | | | |
| AIC | 100 | 133.54 | 3.28 |
| BIC | 99.70 | 97.25 | 2.37 |
| CAIC | 99.70 | 86.25 | 2.09 |
| NEC | 98.48 | 0.50 | 50.42 |
| CLC | 98.48 | 83.39 | 2.06 |
| ICL–BIC | 69.60 | 25.11 | 0.61 |
| Skew 1.5, kurtosis 6: Class-varying variance and covariance parameters (334 of 500 samples)[b] | | | |
| AIC | 100 | 191.03 | 4.70 |
| BIC | 100 | 154.74 | 3.77 |
| CAIC | 100 | 143.74 | 3.49 |
| NEC | 99.10 | 0.63 | 63.49 |
| CLC | 99.10 | 142.75 | 3.52 |
| ICL–BIC | 92.51 | 84.47 | 2.04 |

*Note.* AIC = Akaike's information criterion; BIC = Bayesian information criterion; CAIC = consistent AIC; NEC = normalized entropy criterion; CLC = classification likelihood criterion; ICL–BIC = integrated completed likelihood criterion with BIC approximation.
[a] Mean difference was calculated as Fit1 − Fit2, where Fit1 and Fit2 are the values of the statistic for the one- and two-class models. Percentage change was calculated as (1 − Fit2/Fit1) ∗ 100, where Fit1 and Fit2 are the values of the statistic for the one- and two-class models. Positive values indicate that the fit statistic decreased (improved) by moving to the two-class model. Negative values indicate worse fit of the two-class model relative to the one-class model. [b]The number of samples available for comparison is slightly lower than the number of converged proper two-class solutions reported in Table 1 because the one-class model was also required to converge on a proper solution.

2001; B. O. Muthén & Muthén, 2000; Nagin & Tremblay, 1999). Similarly, support for the two-class model exceeded 98% of replications for the NEC and the CLC fit statistics, with the greatest improvement observed in the skew 1.5, kurtosis 6 condition. These measures thus indicate that the two classes are distinctive and well separated despite the fact that only one class actually exists in the population. Finally, the ICL–BIC, which incorporates penalties for both the number of parameters and poor class separation, was the most conservative of the fit statistics. However, it too supported the two-class model in 70%–99% of replications.

These results supported our first two hypotheses that nonnormality of the repeated measures would be a critical factor influencing both the estimation of a growth mixture model and its fit relative to the correct single-class model. Confronted with the results of the estimated models with nonnormal data, the applied researcher seeking to identify population heterogene-

Table 3
*Relative Fit of One-Class Versus Two-Class Models for Proper Solutions (of 500 Samples, N = 600)*

| Fit statistic | % of time statistic favors two-class model | Mean difference[a] | Mean % change in fit statistic[a] |
|---|---|---|---|
| Skew 0, kurtosis 0: Class-invariant variance and covariance parameters (369 of 500 samples) | | | |
| AIC | 25.93 | −1.47 | −0.01 |
| BIC | 0 | −14.66 | −0.12 |
| CAIC | 0 | −17.66 | −0.14 |
| NEC | 1.90 | −174.83 | −17,482.70 |
| CLC | 1.90 | −402.10 | −3.30 |
| ICL–BIC | 0 | −421.29 | −3.44 |
| Skew 0, kurtosis 0: Class-varying variance and covariance parameters (89 of 500 samples) | | | |
| AIC | 21.35 | −3.63 | −0.03 |
| BIC | 0 | −52.00 | −0.42 |
| CAIC | 0 | −63.00 | −0.51 |
| NEC | 0 | −26.43 | −2,643.08 |
| CLC | 0 | −445.39 | −3.66 |
| ICL–BIC | 0 | −515.76 | −4.21 |
| Skew 1, kurtosis 1: Class-varying variance and covariance parameters (471 of 500 samples) | | | |
| AIC | 100 | 390.64 | 3.20 |
| BIC | 100 | 342.27 | 2.79 |
| CAIC | 100 | 331.27 | 2.70 |
| NEC | 98.73 | 0.41 | 40.54 |
| CLC | 98.73 | 173.34 | 1.42 |
| ICL–BIC | 91.08 | 102.97 | 0.84 |
| Skew 1.5, kurtosis 6: Class-varying variance and covariance parameters (485 of 500 samples) | | | |
| AIC | 100 | 585.62 | 4.80 |
| BIC | 100 | 537.25 | 4.39 |
| CAIC | 100 | 526.25 | 4.29 |
| NEC | 100 | 0.62 | 62.34 |
| CLC | 100 | 389.14 | 3.19 |
| ICL–BIC | 99.18 | 318.77 | 2.60 |

*Note.* AIC = Akaike's information criterion; BIC = Bayesian information criterion; CAIC = consistent AIC; NEC = normalized entropy criterion; CLC = classification likelihood criterion; ICL–BIC = integrated completed likelihood criterion with BIC approximation.
[a] Mean difference was calculated as Fit1 − Fit2, where Fit1 and Fit2 are the values of the statistic for the one- and two-class models. Percentage change was calculated as (1 − Fit2/Fit1) ∗ 100, where Fit1 and Fit2 are the values of the statistic for the one- and two-class models. Positive values indicate that the fit statistic decreased (improved) by moving to the two-class model. Negative values indicate worse fit of the two-class model relative to the one-class model.

ity may make an incorrect inference regarding the number of groups in the population, concluding that at least two groups exist, perhaps more. However, the more appropriate conclusion (given our knowledge of the population model) would be that the superiority of the two-class model directly results from its ability to more accurately represent the nonnormality of the multivariate distribution of the repeated measures when compared with the one-class model.[8] It is important to note that in the latter statement, no inference is made about the number of classes in the population.

[8] We do not claim that a two-class model provides the *best* approximation to the multivariate distribution of the repeated measures. It is possible that three or more components would provide a better fit, potentially exacerbating the problem of identifying nonexistent latent subgroups. The number of latent classes that can be estimated and selected as optimal is likely to depend on a number of factors in addition to the empirical distribution of the data, including the sample size, the number of variables, and the complexity of the model. Because our intent is not to explore the use of these models as a density approximation tool, we do not consider this issue further here.

### *Implications of Overextracting Classes for Parameter Estimates and Standard Errors*

On its surface, this qualification concerning the number of subgroups that exist in the population may appear primarily semantic. However, the importance of this issue becomes more salient when we compare the resulting parameter estimates from one- and two-class models. Because the results from the two sample sizes were nearly identical (with the predictable difference that the estimated standard errors were larger in the $N = 200$ condition), we present only the results from the $N = 600$ condition.[9] Further, we restrict our comparison to the nonnormal conditions, because it was only in these conditions that model fit suggested the presence of multiple classes. Because equality constraints on the variance components were universally rejected in these conditions, we present only the results of the unconstrained model. Finally, we consider only the replications that resulted in proper two-class solutions (representing over 94% of the total replications).

A preliminary hypothesis was that the one-class model would produce consistent parameter estimates and accurate standard errors despite modest violation of the assumption of multivariate normality. As expected, the model recovered the population values of the parameters quite well. The parameter estimates showed little evidence of bias even in the nonnormal conditions. In all cases, the mean relative bias of the parameter estimates was well below the 10% level generally considered acceptable (e.g., Kaplan, 1989).[10] The greatest degree of bias was observed in the variance/covariance components of the model. For instance, $\hat{\psi}_{\alpha}$ was underestimated in both conditions (by averages of 1.4% and 2.2% in the skew 1, kurtosis 1 and skew 1.5, kurtosis 6 replications, respectively), and $\hat{\psi}_{\alpha\beta}$ was overestimated in both conditions (by averages of 1.8% and 2.2% in the skew 1, kurtosis 1 and skew 1.5, kurtosis 6 replications, respectively). The mean relative bias in $\hat{\mu}_{\alpha}$ and $\hat{\mu}_{\beta}$ was under 1% in both conditions. The average estimated standard errors obtained by fitting a one-class model were also quite close to the standard deviations of the sample parameter estimates, indicating that the robust standard errors were unbiased. Overall, the conventional approach to growth modeling, in which only one group is specified, performed well with the mildly nonnormal distributions used in the simulation. Further details on these results are reported in Table 4.

The mean parameter estimates and standard errors for the two-class models are also presented in Table 4. Our interpretation of these estimates depends on whether we have estimated the model to identify population heterogeneity or as a density approximation tool. Assuming our interest is in modeling population heterogeneity, our primary focus would be the within-class estimates. We would interpret them to indicate that there are two distinctive latent subgroups in the population, one with lower initial levels and a slower rate of growth than the other, as shown in Figure 5. Such inferences would in this case be erroneous, because the population model is in fact homogeneous.

Moreover, some relationships that hold in the overall population might not be observed within either class. For instance, in our population model, the correlation between intercepts and slopes was set to a modest positive value ($\rho_{\alpha\beta} = .25$). However, in the two-class models, the value of this correlation within each class was often estimated as negative. The reason is that when the growth mixture model is estimated, the correlational structure of the data is partitioned into a within-class component and a between-class component, where the latter is a function of the mean differences between the classes. In this case, Figure 5 shows that the positive relation between intercepts and slopes in the population was essentially "absorbed" into the differences of the mean trajectories. In almost all cases, the class estimated to have the highest average intercept also had the highest average slope. The remaining relation between intercepts and slopes within classes represents the residual association between these parameters net of the mean differences between the classes. It is thus not surprising that the within-class estimates of this parameter poorly resemble their counterpart in the one-group population model.

It could be argued that estimating the two latent

[9] A full copy of the simulation results can be obtained from Daniel J. Bauer at www4.ncsu.edu/~djbauer or from Patrick J. Curran at www.unc.edu/~curran.

[10] The mean relative bias (MRB) was calculated as

$$\mathrm{MRB} = \frac{1}{R}\sum_{r=1}^{R} 100\left[\frac{(\hat{\theta}_r - \theta)}{\theta}\right],$$

where $R$ is the total number of replications included in the analysis, $\theta$ is a population parameter with a specific value, and $\hat{\theta}_r$ is the estimate of this parameter obtained in replication r.

Table 4
*Population Values for Key Model Parameters Compared With the Mean Value (and Average Standard Errors) of the Parameter Estimates Obtained From One- and Two-Class Models That Converged on Proper Solutions (of 500 Samples, N = 600)*

| Parameter | Population | One-class model | Two-class model[a] | |
|---|---|---|---|---|
| | | | Class 1 | Class 2 |
| | | Skew 1, kurtosis 1 | | |
| | | (500 samples) | (471 samples) | |
| $\mu_\alpha$ | 1.00 | 1.00 (0.05) | 1.50 (0.12) | 0.17 (0.13) |
| $\mu_\beta$ | 0.80 | 0.80 (0.03) | 0.99 (0.06) | 0.51 (0.07) |
| $\psi_\alpha$ | 1.00 | 0.99 (0.13) | 0.79 (0.21) | 0.19 (0.09) |
| $\psi_\beta$ | 0.20 | 0.20 (0.03) | 0.21 (0.06) | 0.05 (0.02) |
| $\psi_{\alpha\beta}$ | 0.11 | 0.11 (0.05) | –0.06 (0.08) | –0.01 (0.03) |
| $CORR_{\alpha\beta}$ | .25 | .26 | –.11 | –.07 |
| % cases | 100 | 100 | 60.7 | 39.3 |
| | | Skew 1.5, kurtosis 6 | | |
| | | (500 samples) | (485 samples) | |
| $\mu_\alpha$ | 1.00 | 1.00 (0.05) | 1.99 (0.22) | 0.65 (0.06) |
| $\mu_\beta$ | 0.80 | 0.80 (0.03) | 1.16 (0.11) | 0.69 (0.03) |
| $\psi_\alpha$ | 1.00 | 0.98 (0.15) | 1.19 (0.50) | 0.40 (0.08) |
| $\psi_\beta$ | 0.20 | 0.20 (0.04) | 0.36 (0.13) | 0.09 (0.02) |
| $\psi_{\alpha\beta}$ | 0.11 | 0.11 (0.05) | –0.15 (0.20) | 0.02 (0.03) |
| $CORR_{\alpha\beta}$ | .25 | .27 | –.19 | .13 |
| % cases | 100 | 100 | 25.3 | 74.7 |

*Note.* CORR = correlation.
[a] Estimated with class-varying variance and covariance parameters.

trajectory classes is a perfectly reasonable way to model the nonnormal aggregate distribution of the data. This argument would be consistent with the use of normal mixture models as a density approximation tool. In this case, the latent trajectory classes could be used to explore the distribution of the repeated measures, with no inferences being made about the existence or characteristics of latent subgroups in the population. For instance, as shown in Figure 5 the degree of departure from the population trajectory is greatest in each condition for the class with the smallest estimated proportion of cases. The separation and asymmetry between the class trajectories can be explained by analogy to the simple univariate mixture discussed earlier. Skewed data require one component to capture the piling up of values at one end of the distribution and another component to capture the long tail at the other end of the distribution. Data that are characterized only by positive kurtosis will generally still require multiple components to account for the peakedness of the distribution, but the means of the two components should be very close. Similarly, it is the degree of skew in the repeated measures that drives the mean trajectories of the two latent classes apart.

In some cases, we can move beyond purely descriptive analyses of the distribution of the repeated measures. If the models estimated within each class have the same form and differ only in their parameter estimates (as is the case here), the within-class estimates can be combined to obtain valid estimates for the aggregate population.[11] In fact, to do so, we can use the same basic equations that Pearson (1894) derived over a century ago. For instance, Equation 9 can be used to calculate the aggregate growth factor means from the within-class means, and Equation 10 can be used to calculate the aggregate factor variances from the within-class variances and means. This is an interesting approach to estimating models with nonnormal data, and further research is needed to evaluate its performance relative to robust or distribution-free methods that do not depend on the use of latent classes (e.g., Bollen, 1996; Browne, 1984; Satorra & Bentler, 1994). However, to date, growth mixture models have rarely been applied for this purpose. In-

[11] We thank Bengt Muthén for this interesting suggestion.

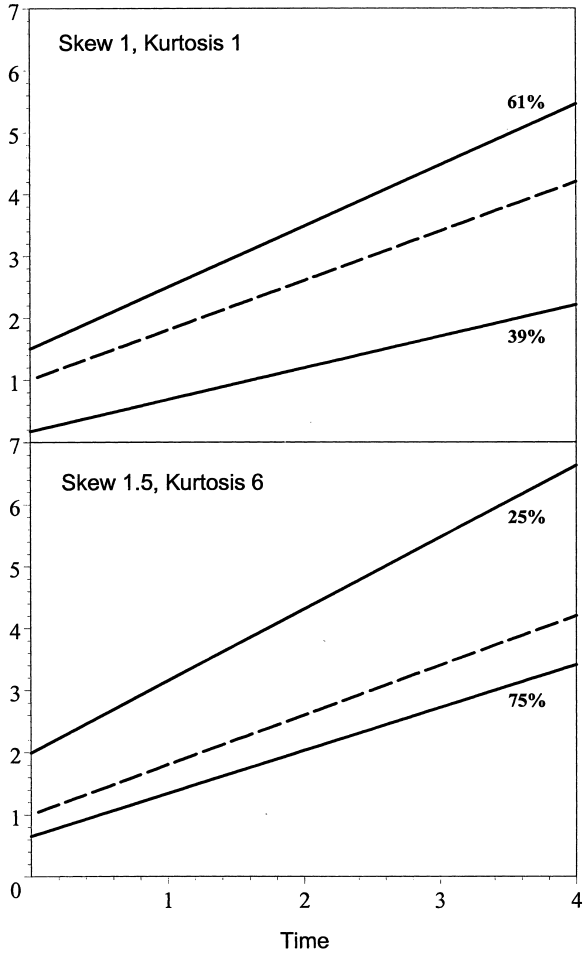


*Figure 5.* The average mean trajectories from convergent two-class models (proper solutions only, $N = 600$), plotted as solid lines. The average estimated percentage of cases belonging to each class is noted beside each trajectory. The single dashed line is the true mean trajectory of the population.

deed, virtually every application of the model of which we are aware has been explicitly motivated by the desire to model population heterogeneity. From this perspective, obtaining valid aggregate estimates is often not of key interest, because the focus is on making valid inferences about the number and characteristics of latent subgroups in the population.

Overall, these results supported our proposed hypothesis that multiple latent trajectory classes would appear optimal for nonnormally distributed data even if these data were generated from a single homogeneous population. In this case, interpreting the trajectory classes to represent the distinctive developmental pathways of unobserved subgroups would be erroneous. Further, if attention is focused on the within-class estimates, it is possible that relationships that exist in the aggregate population will go undetected.

### *Implications of Overextracting Classes When Testing Predictors*

Our final hypothesis was that overextracting classes could also obscure the role of important predictors of individual variability in change over time in the population. To empirically evaluate this prediction, we generated continuous normal data (mean = 0, variance = 10) for a time-invariant covariate that is positively associated with intercepts and negatively associated with slopes using the latent variable model

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \begin{pmatrix} 1.0 \\ 0.8 \end{pmatrix} + \begin{pmatrix} 0.125 \\ -0.03 \end{pmatrix} (x_i) + \begin{pmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \end{pmatrix}, \qquad (13)$$

where the population values of $\mu_\alpha$ and $\mu_\beta$ were unchanged from the unconditional model because the predictor was centered. The values of the regression coefficients relating the predictor to individual intercepts ($\gamma_1 = .125$) and slopes ($\gamma_2 = -.03$) were chosen to represent the kind of moderate effect sizes that are commonly encountered in practice. Comparing the residual variances of the growth factors in the conditional model ($\psi_\alpha = .84$ and $\psi_\beta = .19$) with the variances of these factors in the unconditional model shows that 16% of the variance in the intercept is explained by the predictor, whereas 5% of the variance in the slope is explained by the predictor. Under multivariate normality, the power to detect the effect of $\gamma_1$ would be 0.99 at $N = 200$ and 1.00 at $N = 600$, whereas for $\gamma_2$, the power would be 0.53 at $N = 200$ and 0.94 at $N = 600$ (as calculated using the procedure of Satorra & Saris, 1985).

As noted earlier, there are two ways of treating predictors in growth mixture models. The first approach generalizes Equation 13 and involves estimating class-specific values for $\gamma_1$ and $\gamma_2$. The second approach is unique to the mixture modeling context and treats the covariate as a predictor of individual class membership. Because the earlier results indicated that generally one would not estimate two trajectory classes with normal data, we estimated these models only with the nonnormal data. Further, because across-class equality constraints on the variances and covariance parameters were universally rejected in the nonnormal conditions, we evaluated only the two-class model without such constraints. As

before, only proper solutions were included in the analysis.

*Prediction of interindividual variability within classes.* The first type of conditional model that we examined involves the estimation of class-specific latent variable models.[12] In the present case, we estimated two different two-class models, one in which $\hat{\gamma}_1$ and $\hat{\gamma}_2$ were constrained to be invariant over classes and a second in which they were freely estimated within classes. The latter model reflects a class by covariate interaction. The same estimation method was used as before, including the calculation of robust standard errors. The models were again estimated with six sets of start values, and the same decision algorithm was used to select the solution for analysis.

Comparing the results of fitting one- and two-class conditional models revealed the same trends noted in the unconditional case. That is, the one-class model produced consistent estimates of the population parameters $\gamma_1$ and $\gamma_2$ and accurate robust standard errors. In fact, using an alpha level of .05, we identified $\hat{\gamma}_1$ as significant in 99% of replications at $N = 200$ and in 100% of replications at $N = 600$, whereas $\hat{\gamma}_2$ was significant in 54%–55% of replications at $N = 200$ and in 93%–94% of replications at $N = 600$. These rejection rates are in line with the normal theory-based power estimates, which suggested that $\hat{\gamma}_1$ should be significant in 99% of replications at $N = 200$ and in 100% of replications at $N = 600$ and that $\hat{\gamma}_2$ should be significant in 53% of replications at $N = 200$ and in 94% of replications at $N = 600$.

In contrast, in the two-class models, the within-class parameter estimates roughly resembled the true values of the parameters within the single-group population, but our ability to detect these effects was lower. This decrease in power was due in large measure to the higher standard errors obtained for the within-class estimates, which in turn were a consequence of resolving the total sample into smaller estimated latent classes. Across the two distributional conditions, $\hat{\gamma}_1$ was identified as significant within the larger class at a rate of 87%–88% at $N = 200$ and of 100% at $N = 600$, but within the smaller class it was identified as significant in only 57%–59% of replications at $N = 200$ and in 93%–97% of replications at $N = 600$. Similarly, for the larger class, $\hat{\gamma}_2$ was identified as significant in 42%–49% of replications at $N = 200$ and in 89%–90% of replications at $N = 600$; for the smaller class, $\hat{\gamma}_2$ was significant in only 23%–29% of replications at $N = 200$ and in 55%–61% of replications at $N = 600$. These results show that by artificially rendering a homogeneous population into latent classes, the power to detect the effect of predictors may decrease substantially.

The estimated values of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ also often diverged significantly between the two classes, indicating the presence of a class by predictor interaction. For the skew 1, kurtosis 1 condition, with an an alpha level of .05, likelihood-ratio tests between the constrained and unconstrained two-class models indicated that $\hat{\gamma}_1$ and $\hat{\gamma}_2$ significantly differed between classes in 18% of the replications at $N = 200$ and in 55% of the replications at $N = 600$. Similarly, in the skew 1.5, kurtosis 6 condition, $\hat{\gamma}_1$ and $\hat{\gamma}_2$ differed significantly over classes in 17% of the replications at $N = 200$ and in 47% of the replications at $N = 600$. In actuality, the mean differences between the parameter estimates were quite similar at the two sample sizes, illustrating the fact that as sample size increases, the power to detect effects, even spurious effects, also increases. Additional results from the one-class and two-class unconstrained models at $N = 600$ are provided in Table 5 (results at $N = 200$ were similar and so are not detailed further here).

*Prediction of class membership.* An alternative way to incorporate predictors in the model is to assess their influence on the probability of belonging to particular trajectory classes. Formally, this effect of the predictor is evaluated through the inclusion of a multinomial regression equation in the model for the posterior probabilities. If multiple groups do not exist in the population but are extracted in the sample, treating the covariate as a predictor of class membership may have serious consequences. The situation is loosely analogous to performing a median split on a continuous outcome measure for the purpose of running a logistic regression with a continuous predictor. As J. Cohen (1983) has demonstrated, such practices can lead to the underestimation of effects and often sharply reduce power (see also MacCallum, Zhang, Preacher, & Rucker, 2002). Similarly, in the present case, we expected that treating the exogenous variable as a class predictor within a two-class model would diminish our capacity to detect its effect.

[12] It is important to note that with the inclusion of exogeneous predictors in the model, the distributional assumption shifts to conditional multivariate normality within classes as opposed to unconditional multivariate normality (see Arminger et al., 1999, for further detail on this distinction).

Table 5
*Population Values of Model Parameters Relating a Predictor to the Intercept and Slope Factors Compared With the Mean Value of the Parameter Estimates (and Average Standard Errors) Obtained From One- and Two-Class Models That Converged on Proper Solutions (of 500 Samples, N = 600)*

| Parameter | Population | One-class model | Two-class model[a] | |
|---|---|---|---|---|
| | | | Class 1 | Class 2 |
| | | Skew 1, kurtosis 1 | | |
| | | (499 samples) | (437 samples) | |
| $\gamma_1$ | 0.125 | 0.123 (0.016) | 0.139 (0.025) | 0.070 (0.021) |
| $\gamma_2$ | −0.030 | −0.030 (0.008) | −0.044 (0.013) | −0.021 (0.011) |
| % cases | 100 | 100 | 60.0 | 40.0 |
| | | Skew 1.5, kurtosis 6 | | |
| | | (499 samples) | (458 samples) | |
| $\gamma_1$ | 0.125 | 0.123 (0.017) | 0.189 (0.049) | 0.087 (0.016) |
| $\gamma_2$ | −0.030 | −0.030 (0.008) | −0.061 (0.028) | −0.025 (0.008) |
| % cases | 100 | 100 | 25.5 | 74.5 |

[a] Estimated with class-varying variance and covariance parameters.

The mean logit relating the predictor to class membership was almost identical in all conditions, ranging from 0.10 to 0.11, as was the mean odds-ratio, which ranged from 1.11 to 1.12. The odds-ratios may be interpreted to mean that with each one-unit increment in the predictor, it is 1.11 to 1.12 times more likely that a case belongs to Class 1. For the skew 1, kurtosis 1 condition, this effect was significant in only 30% of the replications at $N = 200$ and in 75% of the replications at $N = 600$. For the skew 1.5, kurtosis 6 condition, the effect of the predictor was significant in only 26% of the replications at $N = 200$ and in 67% of the replications at $N = 600$. These results again supported our hypothesis that overextraction of classes would compromise our ability to detect significant relations between predictors and individual growth.

The estimated logit coefficients and odds-ratios almost always indicated that individuals with high values on the continuous predictor were more likely to belong to the first class. The mean trajectories of the two classes were largely unchanged from those displayed in Figure 5 (based on the unconditional model). That is, Class 1 had both a higher intercept and a more steeply increasing slope than Class 2. As such, one would be tempted to conclude that having a high value on the class predictor increases the likelihood of having both a high intercept and a high slope. However, this conclusion would be erroneous; in the population model, the effect of the predictor is only positive for intercepts and is in fact negative for slopes. Thus, if a predictor has differential effects on the parameters of the growth process, it may be relatively difficult to recover both relationships using the variable as a class predictor. Further, the power to detect the effects of predictors may be diminished if only one group actually exists in the population.

### *Summary*

Taken together, the results of our simulation supported each of our three hypotheses. First, we had hypothesized that it would be difficult to fit a two-class growth mixture model to data drawn from a normal distribution but relatively easy to do so with data drawn from a nonnormal distribution even if the population truly consisted of just a single group. Our empirical results indicated that although two-class models could be fit to data drawn from a normal distribution, it was more difficult than in the nonnormal conditions. Specifically, use of multiple sets of start values was critical to avoid nonconvergence, and it was usually necessary to place equality constraints on the variance and covariance parameters over classes to obtain a proper solution. In direct contrast, when a two-class model was fit to data drawn from nonnormal distributions, the growth mixture model routinely converged on a proper solution even without equality constraints, though in fact no latent subgroups actually existed in the population.

Our second hypothesis was that, under nonnormality, fit statistics would reliably indicate that the two-class model was superior to a single-class model despite the presence of just one group in the population. This hypothesis was supported by the finding that

with normal data, even if two classes could be estimated, the fit of the two-class model was routinely judged to be poorer than that of the correct one-class model. However, when the data were generated from a nonnormal distribution, this was no longer the case. The two-class model was consistently judged to be a better representation of the data than the one-class model despite the fact that only one group was actually present. Finally, our third hypothesis centered on the implications of extracting multiple classes when only one existed. As expected, the overextraction of groups resulted in largely uninterpretable within-class parameter estimates that sometimes obscured relationships present in the aggregate population. It also greatly diminished our ability to detect the full effects of predictors of individual change and frequently led to the identification of spurious class by predictor interactions.

### *Strengths and Limitations*

We believe that this article is characterized by several key strengths. First, we drew on statistical theory to generate a set of specific research hypotheses to be empirically evaluated using computer simulations. Second, we attempted to maximize the external validity of the findings by studying the effects of sample sizes, distributions, and a target model that might be commonly encountered in applied research. Finally, we approached the model estimation procedure from the perspective of an applied practitioner by using multiple sets of start values for each replication and by considering converged and proper solutions. Despite these strengths, there are of course several potential limitations that should be noted.

As with any simulation study, we did not examine all possible experimental conditions. Specifically, additional sample sizes, nonnormal distributions, numbers of repeated measures, and functional forms of growth were not included here. Despite the necessary omission of these additional conditions, we do not believe it presents a significant limitation given the predictable outcomes associated with these variations. For example, it would be expected that smaller or larger sample sizes would predictably influence convergence rates and standard errors. Further, more severe nonnormality would be predicted to increase only the identification of spurious groups. Finally, although many alternative forms of growth could be considered, given that the model is properly specified, we would expect to obtain similar findings but with the added issues associated with the more complex models (e.g., larger numbers of parameters to be estimated). In sum, although additional experimental conditions could be considered, we believe that our findings generalize to a broad set of situations that might be encountered in applied research.

Finally, an important point to highlight is that we made no attempt to analytically or empirically examine the implications of *failing* to extract multiple classes when such heterogeneity does exist in the population. In other work we have demonstrated that it may be quite difficult to identify model misfit when fitting a single-group model to data generated from a multiple-group population (Bauer & Curran, 2001; see also Jedidi et al., 1997). Although this is a critically important situation to consider more closely, the omission of multiple classes has no bearing on the hypotheses tested here. That is, our motivating goal was to explore conditions that might lead to the spurious extraction of multiple classes when only one class truly existed. Much future work is needed to better understand the implications of omitting classes when such classes actually exist in the population.

## Conclusions

Growth mixture models represent an exciting new development that allows applied researchers to examine data in ways not previously possible. Like all finite normal mixture models, growth mixture models may be viewed as having two functions. The function that is most attractive to many social scientists is that these models offer an analytic approach that is consistent with theories emphasizing population heterogeneity in patterns of change over time. The promise of the growth mixture model is to identify taxonomic groups with distinctive trajectories and unique relations to predictors. When the model is used for this purpose, interest centers on the within-class estimates and the population proportions of the latent subgroups. The second function of these models, which is less well known to social scientists, is to provide a means of approximating complex or unknown multivariate distributions of repeated measures. In this case, the component classes are typically viewed as analytic tools for examining the aggregate distribution. There is little expectation that the estimates obtained within classes will actually reflect meaningful population parameters, and they are generally only useful when recombined to examine the characteristics of the aggregate population.

It is important that although these two purposes of

the latent trajectory class model are quite distinct theoretically, each is based on precisely the same analytical model. It may thus be quite difficult to determine analytically which function the estimated latent classes are actually serving. Further, the fit statistics conventionally used to discern the presence of latent subgroups (e.g., the BIC) are also often used to identify the optimal number of components needed to approximate homogeneous but undefined distributions (e.g., Roeder & Wasserman, 1997). Given this fact, it is not surprising that these fit statistics do not distinguish between the two functions of the model. An important implication of this point is that simply identifying the optimal number of latent trajectory classes for the data does not allow strong inferences to be made regarding the number or characteristics of latent subgroups in the population. Although population heterogeneity may well exist, an equally viable explanation is that the trajectory classes simply allow the model to more optimally capture a nonnormal, but ultimately homogeneous, distribution of repeated measures.

Our primary goal in highlighting this issue is to make applied researchers more fully aware of alternative interpretations for their results. In our survey of the literature, including didactic articles by B. O. Muthén (2001), B. O. Muthén (2001), B. O. Muthén and Muthén (2000), and Li et al. (2001), change in model fit with the extraction of additional classes has been interpreted almost universally as a test of population heterogeneity. To our knowledge, the alternative explanation that the additional classes improve the fit of the model only because they serve to approximate an irregular but homogeneous distribution of repeated measures has rarely been considered (an important exception is Nagin, 1999). Similarly, conditional growth mixture models have been recommended to identify subgroups of individuals for whom interventions differ in effectiveness (B. Muthén et al., 2002). However, as our simulation results have demonstrated, with nonnormally distributed data, spurious class by predictor interactions may be obtained even when the population is homogeneous and the predictor has a constant linear effect on the individual trajectory parameters.

Growth mixture models are certainly not unique in this regard; analogous problems beset many other statistical procedures. For example, the broadly used correlation coefficient may be interpreted in multiple ways.[13] A positive correlation between Variables A and B may indicate that A causes B, that B causes A, that a third variable, C, causes both A and B, and that it may be due to the presence of outliers or that it may fail to capture what is really a nonlinear relationship between A and B. The fact that these alternative interpretations exist does not diminish the widespread utility of the statistic. The difference is that the limitation of the correlation coefficient for causal inferences is widely understood, and these alternative interpretations are appreciated and, when possible, empirically evaluated. Our intention is to contribute to a similar level of discourse on the alternative interpretations of growth mixture models.

## Implications for Applied Research

It may be argued that strong substantive theory can be used to guide the interpretation of the model. For instance, if theory indicates that population heterogeneity is likely, then why should we not interpret the latent classes as distinctive population subgroups? The reason is that, in our opinion, this approach reverses the normal hypothetico-deductive process of science. Specifically, using a growth mixture model to test the hypothesis that the population is heterogeneous and then proceeding to interpret the latent classes as true subgroups because that is what theory suggests would be affirming the consequence. The fact that multiple latent classes are optimal for the data no more indicates that the population is heterogeneous than a significant correlation indicates that Variable A causes Variable B. It is important to note that the correlation is *consistent* with the causal process, and the finding that multiple classes are optimal is also *consistent* with the presence of population heterogeneity. As such, testing for the optimal number of components may best be viewed as a method for potentially rejecting (rather than supporting) the hypothesis that the population is heterogeneous. Given the analytical and simulation results we have presented, this may be tantamount to determining that the empirical distribution is nonnormal (given sufficient sample size).

In the case that one is willing to *assume* that this nonnormality reflects a mixture of unobserved groups, each with a multivariate normal distribution, then the growth mixture model is an ideal approach for analyzing the data. However, there are other causes of nonnormality and other approaches for ana-

[13] We thank Andrea Hussong for suggesting this analogy.

lyzing nonnormal data. If one is not willing to assume that the nonnormality reflects population heterogeneity, it may still be possible to gainfully use a growth mixture model. However, in this case, one would be using the model for density approximation, and interpretations should be centered on effect estimates computed for the aggregate population. More conventional approaches for accommodating nonnormal data include robust or distribution-free methods of model estimation (Bollen, 1996; Browne, 1984; Satorra & Bentler, 1994) and the use of nonlinear transformations to normalize the data (e.g., logarithmic transformations). If one is interested in obtaining aggregate estimates from nonnormal data, comparing the results obtained from each of these approaches may be useful.

Another possibility is that there is both a mixture and intrinsic nonnormality or a mixture of nonnormal distributions. This seems especially likely in substantive domains such as drug and alcohol use and antisocial behavior, where growth mixture models have been applied with the most frequency (e.g., Chassin, Pitts, & Prost, 2002; Colder et al., 2001; B. Muthén et al., 2002; B. O. Muthén & Muthén, 2000; Nagin & Tremblay, 1999). Extrapolating from the results presented here, we expect that using a growth mixture model based on a mixture of *normal* distributions would often fail to detect the true number of components, because several normal components might be needed to approximate the true *nonnormal* distribution of each subgroup. An alternative distributional model would, in theory, be preferable (though it might not be analytically tractable). If the correct number of classes is selected, however, Jedidi et al. (1997) have found that the parameter estimates may be consistently estimated despite violation of the assumption of normality within components (though see Arminger et al., 1999, for an exception).

Ultimately, deciding whether the latent classes should best be interpreted as population subgroups may require a programmatic series of research studies, each aimed at testing the validity of the assumption of population heterogeneity. This recommendation follows the classical prescription for assessing construct validity given by Cronbach and Meehl (1955): to construct a nomological network of results that are consistent with the idea of population heterogeneity and that would not necessarily be expected if the population was homogeneous regardless of its distributional form. This task may be difficult. For instance, one might examine the characteristics of the trajectory classes to see if they match theoretical expectations. However, this requires relatively strong predictions about the within-class trajectories. It is our impression that growth mixture models are more typically used in an exploratory mode, with post hoc interpretations of the class trajectories. Another piece of information that would seem to support the notion of population heterogeneity would be if predictors differentially influenced growth in the latent classes. However, as we have shown, spurious class by predictor interactions can be obtained even when the population is truly homogeneous. Further, predictors that bear a significant relationship to the individual growth parameters within a unitary population also tend to distinguish between estimated latent classes. Adequately testing theories of population heterogeneity thus remains both an important empirical and analytical challenge.

## Directions for Future Research

There are several important avenues for future quantitative research on growth mixture models. For instance, the theoretical model we used was based on the simple univariate normal mixture, which differs in several important ways from the growth mixture model. Not only are growth mixture models multivariate, involving a mixture of covariance matrices and mean vectors, but these matrices and vectors are structured. If the structural model is misspecified (i.e., with the incorrect functional form of growth), this might also lead to the overextraction of classes even when the population distribution is multivariate normal. Our reasoning is as follows: In the single-class case, the model-implied means, variances, and covariances will not accurately reproduce the observed moments. A second trajectory class (or more) may be needed to account for this discrepancy (again serving the function of better approximating the observed data). We are currently examining this topic in ongoing research.

It is important, as noted earlier, that we did not consider the case in which population heterogeneity exists but is not explicitly modeled. Early research on this issue has indicated that fitting a conventional one-group model to data arising from a mixture may produce aggregate parameter estimates that fail to reflect the structural relationships among the variables in any one of the individual classes (Jedidi et al., 1997; B. O. Muthén, 1989). Further, traditional model fit statistics may yield little diagnostic information about the mis-

fit of the model (Bauer & Curran, 2001; Jedidi et al., 1997). An important topic for future research will be to determine which error is more severe. That is, in the case that an investigator is uncertain of the number of groups present, is it better to extract groups that may not exist or to employ a conventional model that ignores possible population heterogeneity?

Finally, it will be important to develop and evaluate new fit criteria that may provide a better indication of whether the estimated latent trajectory classes actually reflect population subgroups. The fit statistics we considered are the most widely used but are all based on the manipulation of up to four basic pieces of information: the log-likelihood of the model, the sample size, the number of parameters estimated, and the degree of separation between the latent classes (as detailed in the Appendix). Another approach that has been recommended in the literature is to bootstrap the likelihood ratio test (Aitkin, Anderson, & Hinde, 1981; Arminger et al., 1999; McLachlan, 1987), and more recently, an asymptotically valid likelihood ratio test has been proposed by Lo, Mendell, and Rubin (2001). However, given that these tests depend on the same pieces of information as do the criteria we evaluated, we expect that they would be influenced by nonnormality in a similar way. In reviewing this article, Bengt Muthén suggested that it may be possible to test the plausibility that the latent classes reflect population heterogeneity by evaluating the fit of the model to the higher order moments of the data. We feel that this is an intriguing idea and hope that by drawing greater attention to this issue, successful tests of this sort may be developed in the near future.

## References


Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society Series A (General), 144,* 419–461.

Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika, 64,* 475–494.

Arminger, G., Wittenberg, J., & Schepers, A. (1996). *MECOSA 3 user guide.* Friedrichsdorf/Ts, Germany: ADDITIVE GmbH.

Bauer, D. J., & Curran, P. J. (2001, April). *The impact of model misspecification in clustered and continuous growth modeling.* Poster presented at the biannual meeting of the Society for Research in Child Development, Minneapolis, MN.

Bentler, P. M. (1995). *EQS: Structural equations program manual* (Version 5.0) [Computer software manual]. Los Angeles: BMDP Statistical Software.

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.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 719–725.

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

Biernacki, C., & Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation, 64,* 49–71.

Blåfield, E. (1980). Clustering of observations from finite mixtures with structural information. *Jyvaskyla Studies in Computer Science, Economics, and Statistics, 2.*

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

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika, 61,* 109–121.

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

Brendgen, M., Vitaro, F., Bukowski, W. M., Doyle, A. B., & Markiewicz, D. (2001). Developmental profiles of peer social preference over the course of elementary school: Associations with trajectories of externalizing and internalizing behavior. *Developmental Psychology, 37,* 308–320.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

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

Chassin, L., Pitts, S. C., & Prost, J. (2002). Binge drinking trajectories from adolescence to emerging adulthood in a high-risk sample: Predictors and substance abuse outcomes. *Journal of Consulting and Clinical Psychology, 70,* 67–78.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research, 29,* 468–508.

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, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics, 9,* 15–28.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

Colder, C. R., Mehta, P., Balanda, K., Campbell, R. T., Mayhew, K. P., Stanton, W. R., et al. (2001). Identifying trajectories of adolescent smoking: An application of latent growth mixture modeling. *Health Psychology, 20,* 127–135.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90,* 577–588.

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

Ferguson, T. S. (1983). Bayesian density estimation via mixtures of normal distributions. In M. H. Rizvi, J. S. Rustagi, & D. Siegmund (Eds.), *Recent advances in statistics* (pp. 287–302). New York: Academic Press.

Fleishman, A. I. (1978). A method for simulating nonnormal distributions. *Psychometrika, 43,* 521–532.

Hill, K. G., White, H. R., Chung, I. J., Hawkins, J. D., & Catalano, R. F. (2000). Early adult outcomes of adolescent binge drinking: Person- and variable-centered analyses of binge drinking trajectories. *Alcoholism: Clinical and Experimental Research, 24,* 892–901.

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.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29,* 374–393.

Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research, 24,* 41–57.

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

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

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Catell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). New York: Plenum Press.

McArdle, J. J. (1989). Structural modeling experiments using multiple growth functions. In R. Kanfer, P. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota Symposium on Learning and Individual Differences* (pp. 71–117). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110–133.

McCall, R. B., Appelbaum, M. I., & Hogarty, P. S. (1973). Developmental changes in mental performance. *Monographs of the Society for Research in Child Development, 38* (3, Serial No. 150).

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

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5,* 23–43.

Meredith, W., & Tisak, J. (1984, June). *On "Tuckerizing" curves.* Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100,* 674–701.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

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

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S-T., Yang, C-C., et al. (2002). General growth mixture mod-

eling for randomized preventive interventions. *Biostatistics, 3,* 459–475.

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

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

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

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (Version 2) [Computer software manual]. Los Angeles: Muthén & Muthén.

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

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

Nagin, D., & Tremblay, R. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6,* 18–34.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical modeling* (5th ed.) [Computer software manual]. Richmond: Virginia Commonwealth University, Department of Psychiatry.

Nguyen, T., Muthén, L. K., & Muthén, B. O. (2001). *RUNALL* (Version 1.1) [Computer software]. Retrieved June 12, 2001, from http://www.statmodel.com/runutil.html

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philisophical Transactions of the Royal Society of London. A, 185,* 71–110.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philisophical Transactions of the Royal Society of London. A, 186,* 343–414.

Quandt, R. E., & Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association, 73,* 730–738.

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science, 12,* 103–124.

Raudenbush, S. W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology, 52,* 501–525.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rescorla, L., Mirak, J., & Singh, L. (2000). Vocabulary growth in late talkers: Lexical development from 2;0 to 3;0. *Journal of Child Language, 27,* 293–311.

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. Clogg (Eds.), *Latent variable analysis in developmental research* (pp. 285–305). Newbury Park, CA: Sage.

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

Schulenberg, J., O'Malley, 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, 57,* 289–304.

Sorenson, H. W., & Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica, 7,* 465–479.

Soromenho, G. (1993). Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics, 9,* 65–78.

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

Teicher, H. (1960). On the mixture of distributions. *Annals of Mathematical Statistics, 31,* 55–73.

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

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 448–485). Stanford, CA: Stanford University Press.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika, 48,* 465–471.

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.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer-Verlag.

Vermunt, J. K., & van Dijk, L. A. (2001). A non-parametric random coefficient approach: The latent class regression model. *Multilevel Modelling Newsletter, 13,* 6–13.

White, H. R., Johnson, V., & Buyske, S. (2000). Parental modeling and parental behavior effects on offspring alcohol and cigarette use: A growth curve analysis. *Journal of Substance Abuse, 12,* 287–310.

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.

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

Zucker, R. A. (1986). The four alcoholisms: A developmental account of the etiologic process. In P. C. Rivers (Ed.), *Nebraska Symposium on Motivation: Vol. 34. Alcohol and addictive behaviors* (pp. 27–83). Lincoln: University of Nebraska Press.

## Appendix

## Formulas for the Fit Statistics Used in the Simulation Study

The AIC, BIC, and CAIC are calculated as

$$\text{AIC} = -2 \log L + 2p, \tag{A1}$$

$$\text{BIC} = -2 \log L + p \log N, \tag{A2}$$

$$\text{CAIC} = -2 \log L + p(\log N + 1), \tag{A3}$$

where log $L$ is the estimated log-likelihood of the model, $p$ is the number of parameters in the model, and $N$ is the number of cases in the model. Comparing these equations demonstrates that when log $N > 2$ (or $N > 7$), the BIC will be a more conservative criterion than the AIC for selecting classes. Further, the CAIC can be seen to be more conservative than the BIC by the constant $p$.

The CLC, NEC, and ICL–BIC are derived from the observation that the estimated log-likelihood of the model may be partitioned into two components:

$$\log L = \log L_c + EN(\hat{\tau}). \tag{A4}$$

The first component, log $L_c$, represents the complete-data log-likelihood (or classification likelihood) that would have been obtained had the posterior probabilities of class membership been constrained to values of zero and one (implying perfect classification of observations). The second component, referred to as *entropy,* captures the actual fuzziness of the classification and is calculated as

$$EN(\hat{\boldsymbol{\tau}}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{\tau}_{ik} \log \hat{\tau}_{ik}, \tag{A5}$$

where $\hat{\tau}_{ik}$ is the estimated posterior probability that individual $i$ is a member of group $k$. It can be seen from Equation A4 that as the entropy goes to zero, the mixture likelihood obtains the same value as the classification likelihood, which assumes perfect classification of observations. As such, entropy provides a measure of the quality of classification in which small values indicate a high degree of separation between latent classes.

The CLC and NEC fit statistics are motivated by the desire to select a model that provides optimal classification quality. The CLC is derived directly from Equation A4 as

$$\text{CLC} = -2 \log L + 2EN(\hat{\boldsymbol{\tau}}). \tag{A6}$$

Similarly, the NEC involves a normalization of the entropy component of the likelihood and is given as

$$\text{NEC} = \frac{EN(\hat{\tau})}{\log L - \log L^*}, \tag{A7}$$

where log $L$ is the log-likelihood of the mixture model with $K$ components, and log $L^*$ is the log-likelihood for a single-class model. Strictly speaking, the NEC is undefined when $K = 1$. For comparisons between one-class models and two-class models, Biernacki, Celeux, and Govaert (1999) suggested setting the NEC to 1 for the one-class model. Note that Ramaswamy, DeSarbo, Reibstein, and Robinson (1993) proposed an alternative rescaled entropy measure (ranging from 0 to 1 with 1 indicating perfect classification) which is labeled *entropy* in Mplus 2.01 output (L. K. Muthén & Muthén, 1998). This measure is also undefined for $K = 1$, and no conventions have been established for comparing one- and two-class models using this statistic, so we do not consider it further here.

The ICL–BIC is directly related to the BIC and the CLC, involving penalties for both parameters and poor classification quality. Hence it is a more conservative criterion than either the BIC or the CLC. The ICL–BIC is calculated as

$$\text{ICL–BIC} = -2 \log L + p \log N + 2EN(\hat{\boldsymbol{\tau}}). \tag{A8}$$

Further information on formulas for the fit statistics may be obtained from Bozdogan (1987) and McLachlan and Peel (2000, chap. 6).

Received October 25, 2001
Revision received October 9, 2002
Accepted October 9, 2002 ■