Psychological Methods
2006, Vol. 11, No. 1, 36–53

1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.1.36

Local Solutions in the Estimation of Growth Mixture Models

John R. Hipp and Daniel J. Bauer
University of North Carolina at Chapel Hill

Finite mixture models are well known to have poorly behaved likelihood functions featuring singularities and multiple optima. Growth mixture models may suffer from fewer of these problems, potentially benefiting from the structure imposed on the estimated class means and covariances by the specified growth model. As demonstrated here, however, local solutions may still be problematic. Results from an empirical case study and a small Monte Carlo simulation show that failure to thoroughly consider the possible presence of local optima in the estimation of a growth mixture model can sometimes have serious consequences, possibly leading to adoption of an inferior solution that differs in substantively important ways from the actual maximum likelihood solution. Often, the defaults of current software need to be overridden to thoroughly evaluate the parameter space and obtain confidence that the maximum likelihood solution has in fact been obtained.

Keywords: growth mixture models, start values, latent classes

Many developmental theories posit that different subgroups of individuals follow qualitatively different developmental trajectories over time. Theories of this type are particularly common in the study of developmental psychopathology (e.g., Moffitt, 1993; Schulenberg, O'Malley, Bachman, Wadsworth, & Johnston, 1996), but they have also been advanced in domains as diverse as cognitive and language development (McCall, Appelbaum, & Hogarty, 1973; Rescorla, Mirak, & Singh, 2000) and health and aging (Aldwin, Spiro, Levenson, & Cupertino, 2001). What these theories hold in common is that the hypothesized subpopulations are largely defined by the presentation of behavior over time rather than by grouping variables that are known a priori. Growth mixture models (GMMs) provide an approach for evaluating such theories by identifying latent classes of individuals distinguished by different patterns of change through time.

Mixture models in general are well known to present certain estimation difficulties; namely, there may be many local optima and, in the case of normal mixtures, also singularities on the likelihood surface (singularities are points where the likelihood function goes to infinity, causing model nonconvergence). Estimation of a mixture model with multiple sets of start values is thus often recommended to avoid these irregularities on the likelihood surface and to discriminate local optima from the global optimum (McLachlan & Peel, 2000; Molina, 2000; B. O. Muthén, 2001; B. O. Muthén & Muthén, 2001, p. 373; Solka, Wegman, Priebe, Poston, & Rogers, 1998). In the present article, we examine this issue with respect to the GMM, including the practical impact of these potential problems of estimation on the fitting and interpretation of GMMs in empirical research.

We focus on two variants of the GMM that have appeared often in recent applied research. First, we consider the GMM advanced by B. O. Muthén and Shedden (1999), which generalizes the latent curve approach to analyzing growth trajectories within a structural equation model (McArdle, 1988, 1989; McArdle & Epstein, 1987; Meredith & Tisak, 1984, 1990). This model can also be viewed as either a mixture of random coefficient growth models (Verbeke & Lesaffre, 1996) or as a particular submodel of the structural equation mixture model developed by Jedidi, Jagpal, and DeSarbo (1997a, 1997b) and by Arminger and colleagues (Arminger & Stein, 1999; Arminger, Stein, & Wittenberg, 1999). Its defining feature is the allowance of random effects within classes, that is, within-class heterogeneity in patterns of change. The second specification of the GMM that we consider does not include random effects within classes. This version of the GMM was advanced by Nagin (1999), who referred to it as a *semiparametric group-*

John R. Hipp, Department of Sociology, University of North Carolina at Chapel Hill; Daniel J. Bauer, Department of Psychology, University of North Carolina at Chapel Hill.

We thank the members of the Carolina Structural Equation Modeling group for their helpful comments and input on a draft of this article. The work was partially funded by National Institute on Drug Abuse Grant DA13148 and by the Arts and Science Foundation of the University of North Carolina at Chapel Hill.

Correspondence concerning this article should be addressed to John R. Hipp, University of North Carolina at Chapel Hill, Department of Sociology, Hamilton Hall, CB 3210, Chapel Hill, NC 27599. E-mail: johnhipp@email.unc.edu

based trajectory model. Vermunt and van Dijk (2001) referred to the same model as *latent class regression,* whereas B. O. Muthén (2001) used the term *latent class growth analysis* (LCGA). Despite the proliferation of names for the model, its defining feature is that individuals within a class are assumed to follow precisely the same trajectory (apart from random errors). One advantage of this more restricted specification is that it is simpler to allow for response scales other than the continuous normal (e.g., binary outcomes with a logit or probit link); on the other hand, a disadvantage is that fitting an overly restrictive model can lead to the estimation of spurious classes (e.g., see Bauer & Curran, 2004). In this article, however, we limit our analysis to models that assume conditional normality of the response variables within class.

Although other articles have reported that local optima can occur with GMMs (e.g., B. O. Muthén & Shedden, 1999), the extent of the problem has not previously been explored empirically. In the absence of such studies, recommendations to vary start values are somewhat ambiguous. For instance, little is known about how extensively the parameter space must be probed through variations in start values to locate the true maximum likelihood solution (as opposed to local optima). As of the time of this writing, the default strategies in Mplus 3 (L. K. Muthén & Muthén, 2004) and Latent GOLD 4 (Vermunt & Magidson, 2005) are quite similar: Both programs generate 10 sets of random start values, run through a small number of iterations with each set (10 in Mplus 3 and 50 in Latent GOLD 4), and then take the set with the highest log-likelihood and continue to iterate with that specific set until convergence criteria are satisfied. For the initial iterations, both programs use an expectation–maximization (EM) algorithm to improve stability of estimation; they then switch to a Newton–Raphson, quasi-Newton, or Fisher scoring algorithm to increase the speed of convergence. The PROC TRAJ SAS macro, developed by Nagin and colleagues (Jones, Nagin, & Roeder, 2000), uses a quasi-Newton algorithm and currently has no provision for automatically varying start values, though one can manually input start values. In the more general literature, applications of similarly complex mixture models have reported the use of as many as 5,000 randomized sets of start values (e.g., Dolan, Jansen, & van der Maas, 2004). Although it is unclear whether such a high number is necessary to locate the global solution of a GMM, it is equally unclear whether the much smaller defaults of commonly used software programs are sufficient.

We are also unaware of previous research studying the substantive consequences of failing to locate the global solution for a model. If local solutions are broadly similar to the global solution, mistakenly accepting and interpreting a local solution might not be a serious problem. There is also the issue of whether a researcher estimating GMMs with varying numbers of classes might unknowingly select a final model based on comparing local solutions when a comparison of the global solutions would have suggested the selection of a different model (e.g., one with fewer or more latent classes). In the present article, we take a two-pronged approach to the examination of these issues. First, to highlight the practical importance of these issues, we reanalyze a data set previously presented in a didactic article on the estimation of GMMs. Second, to view how these procedures operate in instances where we know the true model for the data, we present the results of a small-scale Monte Carlo simulation. The goals of the simulation were to provide an initial indication of (a) whether fitting a model with too few or too many latent classes could result in an increased number of local solutions, (b) whether a high number of local solutions could arise because of misspecification of the within-class growth model, and (c) whether estimating a GMM assuming within-class normality with ordinal data could affect the number of local solutions. Together, the case study and simulation allow us to make several recommendations for investigators seeking to use these models in practice. Let us begin, however, by more formally describing the general GMM.

The GMM

Specifying the GMM

The standard latent curve model can be written as

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

where $\mathbf{y}_i$ is a $p \times 1$ vector of observed repeated measures for individual *i,* where p is the number of waves of data, where $\boldsymbol{\eta}$ is a $q \times 1$ vector of latent growth factors (or random coefficients; i.e., intercept and slope), where q is the number of latent growth factors, and where $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of time-specific disturbances that remain after accounting for the underlying latent trajectory. The functional form of the individual trajectories is defined by fixing the coefficients in the $p \times q$ factor-loading matrix $\boldsymbol{\Lambda}$ to predetermined values. For instance, to define a linear latent curve model with latent intercept and slope factors for p equally spaced repeated measures, $\boldsymbol{\Lambda}$ might be set to $(\mathbf{1}\ \mathbf{t})$, where $\mathbf{1}$ is a column of ones and $\mathbf{t} = (0, 1, 2, \ldots, p - 1)'$.

It is traditionally assumed that the growth factors and time-specific residuals are mutually independent and multivariate normally distributed as

$$\begin{bmatrix} \boldsymbol{\eta}_i \\ \boldsymbol{\varepsilon}_i \end{bmatrix} \approx N\left(\begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Theta} \end{bmatrix}\right), \tag{2}$$

where α is a $q \times 1$ vector of growth factor means, $\boldsymbol{\Psi}$ is the $q \times q$ covariance matrix of the growth factors, and $\boldsymbol{\Theta}$ is the $p \times p$ covariance matrix of the time-specific residuals (typically constrained to be a diagonal matrix). Because the repeated measures are expressed as a linear combination of

the normally distributed η_i and ε_i, $\mathbf{y}_i$ is then also multivariate normally distributed with probability density function

$$f(\mathbf{y}_i) = \phi[\mathbf{y}_i;\boldsymbol{\mu}(\boldsymbol{\theta}),\boldsymbol{\Sigma}(\boldsymbol{\theta})], \quad (3)$$

where ϕ designates a multivariate normal probability density function for $\mathbf{y}_i$ with a $p \times 1$ model-implied mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ and a $p \times p$ model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ given by

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\theta}) &= \boldsymbol{\Lambda}\boldsymbol{\alpha} \\ \boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \end{aligned} \quad (4)$$

and $\boldsymbol{\theta}$ is the vector of parameters from all model matrices. Maximum likelihood estimation then seeks estimates $\hat{\boldsymbol{\theta}}$ that maximize the likelihood that the observed data vectors $\mathbf{y}_i$ would have been drawn from a multivariate normal distribution with mean vector $\boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$ and covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.

The GMM of B. O. Muthén (2001) extends this traditional latent curve model by permitting the estimation of K latent classes each with its own latent curve model. The probability density function for the GMM is thus a finite mixture of normal distributions of the form

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k[\mathbf{y}_i;\boldsymbol{\mu}_k(\boldsymbol{\theta}_k),\boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k)], \quad (5)$$

where π_k is the unconditional probability that an observation will be drawn from latent class k (also interpretable as the proportion of cases belonging to class k), and ϕ_k now represents the multivariate normal probability density function for latent class k. The model-implied mean vector and covariance matrix of a latent curve model again govern each class distribution:

$$\begin{aligned} \boldsymbol{\mu}_k(\boldsymbol{\theta}_k) &= \boldsymbol{\Lambda}_k\boldsymbol{\alpha}_k \\ \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) &= \boldsymbol{\Lambda}_k\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k. \end{aligned} \quad (6)$$

Each model matrix has been subscripted by k to indicate that the parameters within the matrix can potentially vary over classes.

In practice, it is common to assume that each group follows the same basic functional form of growth, such that the factor-loading matrices can be constrained to be invariant over classes (i.e., $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}$ for all k). The growth factor covariance matrices and residual covariance matrices are also often assumed to be invariant over classes (i.e., $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}$ for all k). As can be seen from Equation 6, in combination, these constraints imply that the model-implied covariance matrices of the latent classes are equal (or $\boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ for all k) and that the only differences between classes are in the model-implied means of the repeated measures as determined by the class-varying growth factor means $\boldsymbol{\alpha}_k$. An advantage of making the within-class covariance matrices invariant, as we describe more fully in the next section, *The Problem of Local Solutions,* is that it ensures the absence of singularities, ensures the existence of a global solution, and potentially reduces the number of local solutions. Additionally, if the latent growth factors are assumed to be fixed (not random) coefficients within each class, then the within-class covariance matrix for the latent factors can be set to zero (i.e., $\boldsymbol{\Psi}_k = \mathbf{0}$) and the LCGA model of Nagin (1999), Vermunt and van Dijk (2001), and B. O. Muthén (2001) is obtained.

The Problem of Local Solutions

Unlike most latent curve models (for homogeneous populations), the optimization of GMMs can be quite difficult. For instance, the estimation of latent curve models by maximum likelihood is typically insensitive to the start values used for the parameter estimates in the initial iteration. As such, users usually do not concern themselves much with evaluating whether the solution they have obtained is in fact the global maximum or just a local (inferior) solution because it is almost always the former. With GMMs, however, the situation is not as simple. Like many other methods that produce clusters or classes, such as K-Means cluster analysis, latent class analysis, and finite mixture modeling more generally (McLachlan & Peel, 2000; Steinley, 2003), GMMs are susceptible to local solutions (B. O. Muthén & Shedden, 1999). An intuitive explanation for this common problem is that it seems that some combinations of parameter estimates fit local features of the data well (e.g., particular bumps in the distribution), but do not fit the overall features of the data as well as other values for the same estimates (e.g., the general shape of the distribution).

For normal mixture models, imposing constraints on the covariance matrices of the latent classes has particularly important implications for the estimation of the model. In the absence of specific constraints on these matrices (e.g., equality or proportionality), the likelihood surface will contain many singularities—spikes going to infinity when the mean vector of a class becomes centered on the observed values of a single case and the variance (or determinant) of that class goes to zero (Kiefer & Wolfowitz, 1956; B. O. Muthén & Shedden, 1999).[1] This would seem to present

[1] There is reason to believe that such singularities will occur less commonly for GMMs than for unstructured normal mixtures, given the time trend imposed on the mean vector. For the mean vector of a class to become centered on a single observation, the data vector for that observation would have to follow exactly the functional form of change over time that is imposed on the class mean vector. For instance, choosing the first observation arbitrarily, the equality $\boldsymbol{\mu}_k(\boldsymbol{\theta}_k)=\mathbf{y}_1$ can only be achieved if $\mathbf{y}_1$ follows exactly the time trend imposed on $\boldsymbol{\mu}_k$ (e.g., linear or quadratic). Although it is certainly possible that one or a subset of observations will meet this condition, not all observations will do so, reducing the number of singularities.

quite a problem; however, because these singularities occur on the edges of the parameter space, they can often be avoided by judiciously selecting starting values on the interior of the parameter space (Biernacki & Chretien, 2003). If the maximum likelihood estimator wanders into one of these singularities, it will simply fail to converge, and the model can be reestimated from a different starting position. The more insidious problem is that one can obtain spurious maxima that border on these singularities for small but tightly clustered sets of data points (Biernacki, 2003; Day, 1969; McLachlan & Peel, 2000). One is then left with the difficulty of ascertaining which of a number of possibly spurious solutions is best, where in this specific context no global solution actually exists (given that the likelihood is unbounded).

Placing invariance constraints on the class covariance matrices rids the likelihood surface of these singularities. As noted above, in a GMM, placing invariance constraints on $\mathbf{\Lambda}_k$, $\mathbf{\Psi}_k$, and $\mathbf{\Theta}_k$ together implies that the within-class covariance matrices are also invariant, thereby ensuring the absence of singularities and potentially reducing the number of local solutions. Moreover, there is the comfort of knowing that a global solution exists to be found. However, problems remain: These invariance constraints may not be consistent with theory, and they can have a large impact on model estimates (Magidson & Vermunt, 2004; McLachlan & Peel, 2000). Further, multiple optima may still be present even with these constraints in place. There is thus no clear consensus regarding whether these constraints should be applied. Nevertheless, in our studies, we investigated only models imposing equality constraints on the within-class covariance matrices, expecting that whatever issues we identified would only worsen when such constraints were not present. We now turn to our case study analysis, followed by a small simulation study on the optimization of GMMs.

A Case Study on Local Optima in GMMs

Our case study reexamined an empirical example presented by B. O. Muthén and Muthén (2000) within the context of a didactic article on GMMs. The data came from the National Longitudinal Survey of Youth, a nationally representative multistage probability sample of 12,686 males and females born between 1957 and 1964, with an oversampling of African Americans, Latinos, and economically disadvantaged White youths. B. O. Muthén and Muthén focused on individuals born in 1964, resulting in 1,192 observations; listwise deletion yielded a final sample of 924. The questions used in the analysis concerned heavy drinking behavior and were collected in 1982, 1983, 1984, 1988, 1989, and 1994. The question read, "How often have you had 6 or more drinks on one occasion during the last 30 days?" The responses were recorded as follows: 0 = *never,* 1 = *once,* 2 = *2 or 3 times,* 3 = *4 or 5 times,* 4 = *6 or 7 times,* 5 = *8 or 9 times,* 6 = *10 or more times.*

As with any analysis, some of the modeling decisions made in the original analyses could be debated either theoretically or on statistical grounds. For instance, probability weights were not used, partially missing data were not included, and the outcome variable was treated as a continuous variable despite being ordinal in nature (perhaps under an implicit assumption that the scale is linear in terms of *intensity* of heavy drinking).[2] This simplification of the analysis is understandable given the didactic goals of the original article, and our purpose was not to reevaluate these decisions here. In an effort to retain focus on the specific issue of local solutions, we adopted all of the modeling decisions made by B. O. Muthén and Muthén (2000) in their analyses, with the exclusive exception that we used an extensive range of starting values in the estimation of the models. This approach allowed us to contrast the results we obtained here when exploring a range of start values with the published results of the primary authors. Our adoption of the authors' procedures extended to the criteria used in model selection, particularly the final number of latent classes. Like the authors, we focused on the comparative fit and substantive content of models estimated with successively more classes. A penalized likelihood criterion, such as Bayes's information criterion (BIC), is typically used to assess comparative fit. Although the BIC is but one possible measure by which to compare GMMs, its performance can be regarded as emblematic of other related criteria.

Models Estimated

Although many models were estimated in B. O. Muthén and Muthén's (2000) original article, we have chosen to focus on just the core subset of those models for reanalysis here. One reason for our selection of B. O. Muthén and Muthén's article for reanalysis is that the authors presented results from GMMs both with and without random effects. In their presentation, they reserved the acronym GMM exclusively for models that included random effects within classes, using the acronym LCGA to refer to models without random effects. We follow the same convention in this article. More specifically, the models estimated by B. O. Muthén and Muthén were of the form given in Equations 5 and 6, with model matrices specified by a quadratic growth model (where the time axis was nonlinearly transformed to permit asymmetric curvature in the trajectories):

[2] This latter decision to treat the ordinal measures as continuous also violates the assumption of these models that the data are normally distributed within classes. Nevertheless, this approach is common to many applied analyses and even other didactic articles (see Li, Duncan, Duncan, & Acock, 2001). We thus consider the potential implications of this decision for the optimization of the model in the *Misspecifying the Distribution of the Data* section.

$$\boldsymbol{\Lambda}_k = \begin{bmatrix} 1 & -3.008 & 9.048 \\ 1 & -2.197 & 4.827 \\ 1 & -1.621 & 2.628 \\ 1 & -0.235 & 0.055 \\ 1 & 0 & 0 \\ 1 & 0.884 & 0.781 \end{bmatrix}, \quad \boldsymbol{\Theta}_k = DIAG \begin{bmatrix} \theta_{11} \\ \theta_{22} \\ \theta_{33} \\ \theta_{44} \\ \theta_{55} \\ \theta_{66} \end{bmatrix},$$

$$\boldsymbol{\alpha}_k = \begin{bmatrix} \alpha_{1k} \\ \alpha_{2k} \\ \alpha_{3k} \end{bmatrix}, \quad \boldsymbol{\Psi}_k = \begin{bmatrix} \psi_{11} & & \\ \psi_{21} & \psi_{22} & \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix} or \; \boldsymbol{\Psi}_k = \mathbf{0}. \quad (7)$$

Note that only the means of the latent growth factors were allowed to vary over classes (given that only the parameters of the $\boldsymbol{\alpha}$ vector contain a k subscript) and that the covariance matrices of the latent growth factors were either null for the LCGA models or were estimated but held equal over classes for the GMMs. In both cases, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Theta}_k$, and $\boldsymbol{\Psi}_k$ were all constrained to be invariant, implying that the model-implied covariance matrices for the repeated measures were also equal over classes. As such, the likelihood for each fitted model is free of singularities and a global maximum exists.

Summary of Results Reported by B. O. Muthén and Muthén (2000)

In the original analyses, B. O. Muthén and Muthén (2000) found the nine-class LCGA solution to have the optimal BIC value. However, they ultimately selected the four-class GMM for interpretation given that it captured information similar to that captured by the nine-class LCGA without the need for additional classes (the allowance of random effects in a GMM typically reduces the number of classes necessary to model individual differences in change over time relative to an LCGA). Although the authors also found superior BIC values for GMMs with more classes, they pointed out that these other models contained little information about substantive classes beyond that obtained in the four-class model. For the final four-class model they selected, the most prevalent class (73% of the sample) was characterized by little heavy drinking at any point between ages 18 and 30. The second class (15%) showed modestly high levels of heavy drinking at age 18 with gradual declines by age 30. The third and smallest latent class (5%) engaged in repeated heavy drinking at age 18 but rapidly desisted to the same levels as the second latent class. Finally, the fourth latent class (7%) was of particular substantive interest because it showed little heavy drinking at age 18 but showed a rapid increase in heavy drinking from that point forward. We now explore whether different conclusions might have been reached by extensively evaluating the sensitivity of the estimated models to different starting values.

Estimation and Variation of Start Values

All models were estimated with Mplus 3.01 using the default accelerated EM algorithm described earlier. We generated 999 randomized sets of start values for each model that was estimated. Although Mplus (beginning with Version 3) provides an internal facility for randomizing start values, we preferred to select start values using our own procedures (a point to be discussed further in the Software Considerations section).[3] Because there is little guidance in the literature regarding proper ranges of start values, we began by first specifying what we regarded as a wide but reasonable range of possible values for each parameter of interest. Selecting too narrow a range could decrease the possibility of finding the global solution, while selecting too wide a range yields more implausible combinations and likely leads to more nonconverged solutions. Therefore, we adopted the following strategy.

Start values for growth factor means. Following a common diagnostic strategy, we estimated an ordinary least squares quadratic regression model for each case individually (Bollen & Curran, 2005; Singer & Willett, 2003). We then used the first and third quartiles of the individual ordinary least squares estimates of each trajectory parameter as the minimum and maximum start values for the GMMs and LCGA models.

Start values for growth factor variances. We chose a small value (.05) as the minimum for the range of start values. For the maximum value, we estimated a latent growth curve model (or one-class GMM). The estimated variances of the three latent growth factors from this model were used as the upper bounds for start values for these parameters.

Start values for growth factor covariances. We constrained the covariances between the latent growth factors to have correlations ranging between −.75 and .75.

Start values for residual variances. We computed the observed variance for each outcome variable in the raw data and used 20% and 80% of this variance as the minimum and maximum values.

Start values for class probabilities. In Mplus, the class probabilities are modeled indirectly by a multinomial regression specification. Namely,

$$\pi_{ik} = \frac{\exp(\alpha_{c_k})}{\sum_{k=1}^{K} \exp(\alpha_{c_k})}, \quad (8)$$

[3] Our programs can be obtained from John R. Hipp upon request. While admittedly not as user friendly as commercial software, these programs have two key advantages: The user can input a different start value range for each parameter, and the results from each random start are stored in an external file. A limitation of this software is that it is designed only to interface with Mplus (Versions 2 or 3).

where the scalar α_{c_k} is the intercept for class k, and the coefficient for the last class (the reference class) is standardized to zero for identification ($\alpha_{c_k} = 0$). We varied the intercepts α_{c_k} between −3 and 3, allowing for the possibility of small classes.

Within these ranges, start values were determined by random draws from a uniform distribution. Table 1 lists the range of start values we used for each parameter in these initial models. Except for parameters constrained to be invariant over classes, unique start values within the specified range were generated for each parameter in each class.

Number of Unique Solutions, Frequency of the Optimal Solution, and Rate of Convergence

We defined the number of unique solutions for a given model as the number of different log-likelihood values, or optima, obtained by fitting that same model to the data from different starting positions. For our case study, more unique solutions were obtained for models including more latent classes, regardless of whether random variability was permitted within classes. This effect was particularly pronounced for the LCGA model, as shown in Table 2. For instance, for the four-class model, 3 unique maxima were detected. For six classes, the number of unique solutions jumped to 30, and this value increased for models with more classes.

The pattern was similar for the GMM, as unique solutions were found more frequently in models with increasing numbers of classes. When multiple solutions were found, regardless of the fitted models, the highest optima were not the most frequently obtained solutions. Indeed, this finding may explain why B. O. Muthén and Muthén (2000) presented the four-class solution with the second highest log-likelihood as the ideal model for the data despite the fact that an alternative four-class solution had a higher log-likelihood. Either B. O. Muthén and Muthén (2000) did not detect this alternative solution (which would not be surprising given that it was obtained with only 3% of the converged random starts), or they in fact detected this solution but decided that it was spurious, a possibility we evaluate in the *Substantive Differences Between Solutions* section. Similar patterns were observed for models with more latent classes, as seen in Table 3. Tables 2 and 3 also show that the number of random starts leading to model convergence declined monotonically as successively more latent classes were added to the models.[4] The combination of fewer converged solutions along with a larger number of unique solutions implies that different solutions are frequently found for models with many classes. To illustrate this, we plot the ratio of unique solutions to converged solutions in Figure 1. As can be seen, for the five-class GMM, every fifth random start resulted in a new solution (a ratio of about .2), whereas for the six-class GMM, every other set of start values converged on a new solution (a ratio of about .5).[5]

Table 1
Range of Start Values for the National Longitudinal Survey of Youth Data and the Simulation

Parameter	Start value range	
	Lower bound	Upper bound
α_{c_k} [a]	−3.00	3.00
α_{1k}	0.00	1.51
α_{2k}	−0.08	0.19
α_{3k}	0.05	0.40
θ_{11k} [b]	0.82	3.29
θ_{22k} [b]	0.67	2.69
θ_{33k} [b]	0.58	2.30
θ_{44k} [b]	0.47	1.90
θ_{55k} [b]	0.51	2.05
θ_{66k} [b]	0.68	2.70
ψ_{11k} [b]	0.05[d]	1.10
ψ_{22k} [b]	0.05[d]	0.68
ψ_{33k} [b]	0.05[d]	0.08
ρ_{21k} [b,c]	−0.75[d]	0.75
ρ_{31k} [b,c]	−0.75[d]	0.75
ρ_{32k} [b,c]	−0.75[d]	0.75

Note. Unique start values were generated for each class from a random uniform distribution with the specified range unless otherwise noted.
[a] α_{c_k} is the intercept of the multinomial logit in Equation 8 and is a direct nonlinear function of π_k. [b] These parameters were held invariant over latent classes, thus only one start value was generated for all classes. [c] To avoid out-of-bounds values for the covariance parameters, ψ_{21k}, ψ_{31k}, ψ_{32k}, a random value was generated within the specified range for the corresponding correlation (i.e., ρ_{21k}) and was then transformed to a value for ψ_{21k} utilizing the start values for ψ_{11k} and ψ_{22k} by the following formula: $\rho_{21k} \times \sqrt{\psi_{11k}\psi_{22k}}$. [d] The variances and covariances of the latent factors were constrained to zero in the latent class growth analysis models.

Model Fit Comparisons

To compare models with different numbers of latent classes, we considered the range of BIC values obtained across solutions. The results for the LCGA model are presented in Figure 2, while the same results for the GMM model are presented in Figure 3. In both of these figures, we see considerable overlap between adjacent models, with the overlap increasing as the number of classes increases. These figures also make clear that comparison of BICs from inferior solutions across models could suggest adoption of a different model than comparison of the minimum BICs (e.g., global solutions). The BIC values reported by B. O.

[4] We did not find that convergence failed because the number of iterations was exceeded. Instead, in most instances nonconvergence occurred because the algorithm was iterating toward a degenerate solution in which one or more classes contained zero cases.

[5] Note that the absolute value of this ratio depends in large part on the number of starts and the range of start values used. Relative comparisons across models are more useful, given that the number of starts and the range of start values was held constant.

Table 2
Results Obtained by Fitting Latent Class Growth Analysis Models (No Random Effects Within Classes) With 999 Sets of Start Values to National Longitudinal Survey of Youth Data

Class[a]	No. of estimated parameters[b]	Converged models (%)[c]	No. of unique solutions[d]	Best converged solutions (%)[e]	Modal converged solutions (%)[f]
1	9	100.0	1	100.0	
2	13	100.0	1	100.0	
3	17	99.9	1	100.0	
4	21	86.0	3	12.1	83.9
5	25	78.8	6	39.8	
6	29	72.4	30	1.0	18.8
7	33	61.2	37	5.2	23.4
8	37	49.6	55	14.5	19.6
9	41	36.2	81	0.6	14.9
10	45	25.4	81	0.8	15.0

Note. Models were estimated with invariant $\mathbf{\Theta}_k$ and $\mathbf{\Psi}_k = \mathbf{0}$.
[a] Number of classes estimated for the model. [b] Number of parameters estimated for the model. [c] Percentage of models that converged out of 999 randomized starting values. [d] Number of different likelihood values found for a model with a given number of classes. [e] Percentage of converged solutions that resulted in the maximum likelihood value. [f] Percentage of converged solutions that resulted in the most frequent likelihood value, indicated only when the modal solution and best solutions (as defined by log-likelihood) differed.

Muthén and Muthén (2000) are also shown in these figures and, in general, do not correspond to the best solutions, but rather reflect local optima. This was particularly the case for models with many optima. This pattern of findings strongly suggests that failing to extensively explore the likelihood surface of a given model for multiple optima can limit both the ability to detect the maximum likelihood solution (particularly as model complexity increases with more classes or the inclusion of random effects) and the ability to select the best model from a set of competing models.

Substantive Differences Between Solutions

We now consider the solutions chosen in the original analysis and whether selection of a local solution might lead to substantively different conclusions relative to those that

Table 3
Results Obtained by Fitting Growth Mixture Models With 999 Sets of Start Values to National Longitudinal Survey of Youth Data

Class[a]	No. of estimated parameters[b]	Converged models (%)[c]	No. of unique solutions[d]	Best converged solutions (%)[e]	Modal converged solutions (%)[f]
1	15	100.0	1	100.0	
2	19	99.6	2	53.5	
3	23	56.6	5	21.1	63.2
4	27	20.2	11	3.2	46.0
5	31	12.0	22	0.9	19.5
6	35	4.8	24	2.2	10.9
7	39	4.4	26	2.4	11.9
8	43	2.2	18	4.8	14.3
9	47	1.8	17	5.9	

Note. The growth mixture models were fitted allowing random effects within classes. Models were estimated with invariant $\mathbf{\Theta}_k$ and $\mathbf{\Psi}_k$.
[a] Number of classes estimated for the model. [b] Number of parameters estimated for the model. [c] Percentage of models that converged out of 999 randomized starting values. [d] Number of different likelihood values found for a model with a given number of classes. [e] Percentage of converged solutions that resulted in the maximum likelihood value. [f]Percentage of converged solutions that resulted in the most frequent likelihood value, indicated only when the modal solution and best solutions (as defined by log-likelihood) differed.

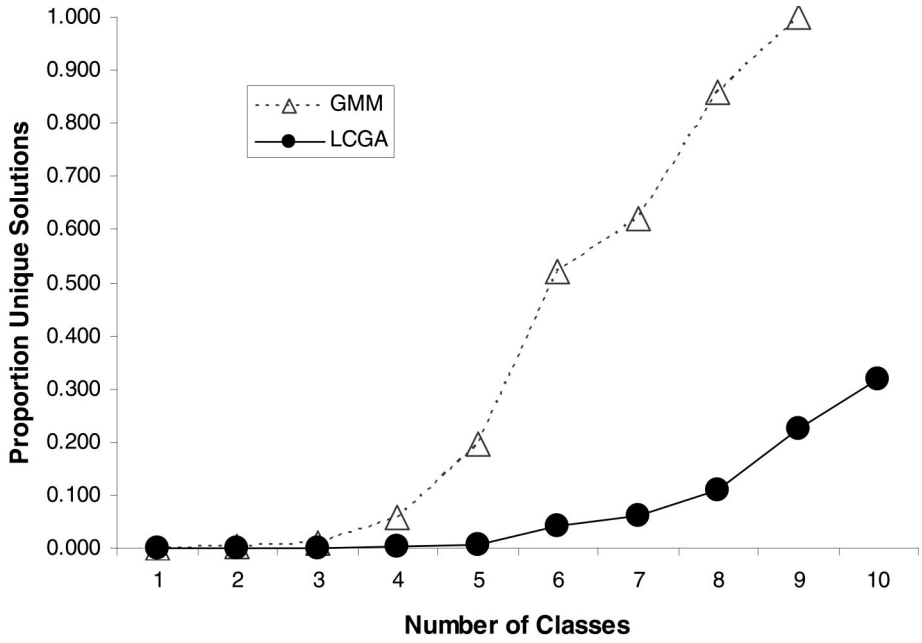

Figure 1. The ratio of unique solutions to converged solutions found by estimating latent class growth analysis (LCGA) and growth mixture models (GMMs) to the National Longitudinal Survey of Youth data from different randomized start values increases with the number of classes estimated.

would be drawn from the global solution. For the LCGA solutions, the difference in the optimal solution and the most frequent solution were minimal—the class trajectories were broadly similar regardless of which solution was selected. However, the solutions for the four-class GMM model showed more pronounced differences. Figure 4 contrasts two particularly salient four-class solutions. Figure 4A shows the trajectories for the four classes found in the solution with the highest log-likelihood, arrived at by only about 3% of our converged random starts. Contrasted with this is the second best solution, shown in Figure 4B, which is the solution presented by B. O. Muthén and Muthén

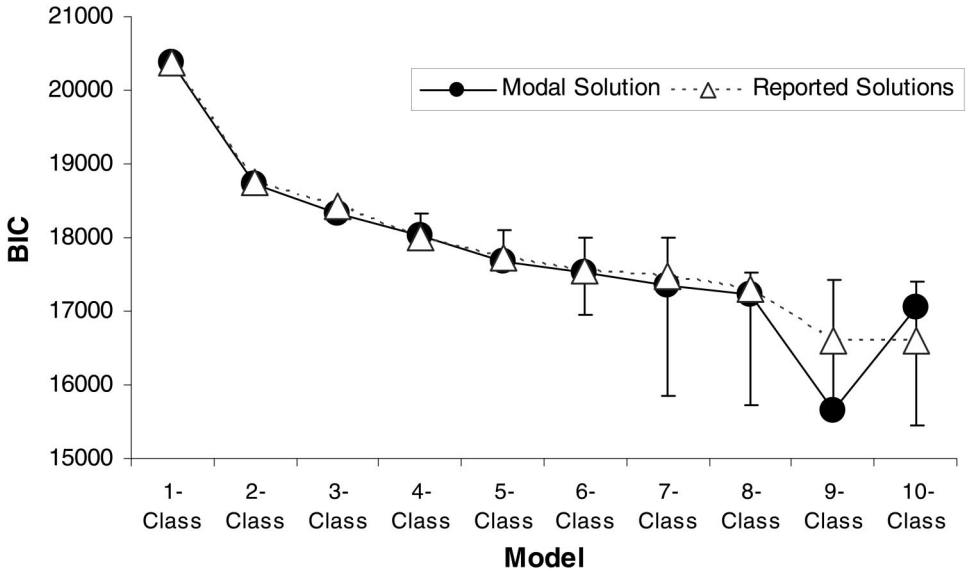

Figure 2. The minimum, maximum, and modal Bayes's information criterion (BIC) values obtained from different optima for the same model, plotted for latent class growth analysis models with between 1 and 10 latent classes fit to the National Longitudinal Survey of Youth data.

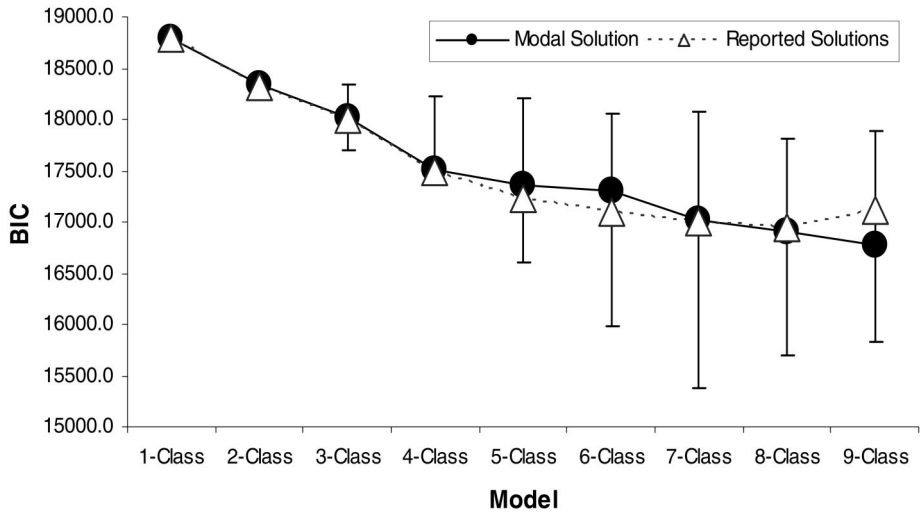

Figure 3. The minimum, maximum, and modal Bayes's information criterion (BIC) values obtained from different optima for the same model, plotted for growth mixture models with between one and nine latent classes fit to the National Longitudinal Survey of Youth data.

(2000) and was obtained by 46% of our converged random starts. Although this second solution contains a latent class with rapidly increasing heavy drinking that Muthén and Muthén noted as particularly important theoretically, no latent classes show this pattern in the solution with the highest log-likelihood. The substantive implications of the two solutions are thus markedly different.

We considered two possible resolutions to this problem. First, given that the classes in question are rather small, we considered whether the four-class model might be overfitting the data, such that the results of a three-class model might be more stable. A second possibility, suggested by B. O. Muthén and Shedden (1999), is that the identification of alternative solutions may signal that too few classes have been estimated. That is, the classes plotted in Figure 4 might be subsets of a larger group of latent classes. To better evaluate these possibilities, we also considered the results of the three- and five-class models. Rather than resolving the problem, a similar story emerged in each case: In each instance, the solution with the highest log-likelihood did not contain the increasing trajectory class, though this class did appear in solutions with inferior log-likelihood values. Although we are not in a position to adjudicate between these two possible models—one containing an upward sloping trajectory and one instead containing an intermediate one—our analyses illustrate how local solutions can lead to different substantive conclusions.

Summary

In this case study, many local solutions were detected despite invariance constraints on the class covariance matrices. Moreover, the solution with the highest log-likelihood for a given model was often obtained infrequently and was rarely reported in the original analysis. Typically, only a small percentage of our random starts led to convergence on the solution with the highest likelihood for an estimated model. Moreover, different solutions for the same model sometimes differed in substantively important ways. The most dramatic finding was that the increasing-alcohol-use class of theoretical interest in the original analysis was not present in the four-class solution with the highest log-likelihood. More optimistically, although the smallest latent classes showed little stability across solutions, the more predominant patterns (e.g., the large class showing little heavy drinking at any age) emerged consistently.

While our case study illustrates how sensitive these models are to starting values when using maximum likelihood with the EM algorithm, a limitation is that we do not know the true population model. Hence, we cannot definitively determine which solution is the correct one for the data. To address this limitation, we next present the results of a small Monte Carlo simulation designed to evaluate to what extent local solutions may detract from our ability to recover the true model for the data.

Monte Carlo Simulation

For the simulation, we focused on two population models. Our first population model was a four-class GMM using the estimates reported by B. O. Muthén and Muthén (2000) as population parameters. For our second population model, we generated data for an LCGA: Because most applications of this model find three to six classes, the four-class population model we used to generate the data can be considered

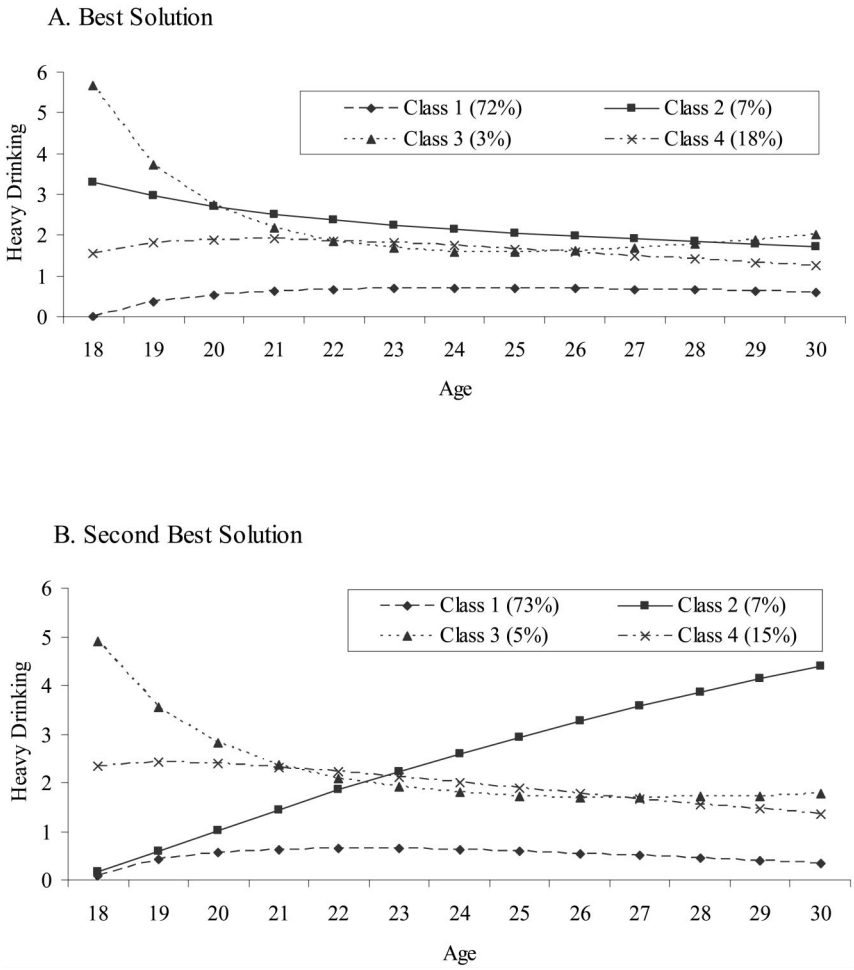

Figure 4. The model-implied mean trajectories of the four-class growth mixture models fit to the National Longitudinal Survey of Youth data. (A): The trajectories implied by the best fitting solution. (B): The trajectories implied by the second-best fitting solution (as presented by B. O. Muthén & Muthén, 2000).

representative of empirical studies using this technique. The population-generating model was in this case identical to the GMM population model, with the exception that the covariance matrix of the latent variables was specified as a null matrix. We include a complete description of our data-generation procedures in the Appendix.

For each of our two population models, we simulated 200 data sets with 924 observations each (the sample size of our case study). Although the data for our case study were ordinal, we simulated continuous data here unless otherwise noted. We focused on three questions in our simulation. First, we asked whether fitting a model with too few or too many latent classes could result in an increased number of local solutions. For each data set, we estimated three models, one model with the correct number of classes—four—and then models with one too few and one too many classes. Second, we asked whether a high number of local solutions could arise because of misspecification of the within-class growth model. For each data set we estimated four-class models with (a) a linear within-class specification with the transformed time codes used in the original analysis (i.e., omitting the quadratic effect), or (b) a quadratic within-class specification in which the time codes were not nonlinearly transformed (i.e., failing to permit asymmetric curvature). Third, we briefly considered whether estimating a GMM assuming within-class normality with ordinal data could affect the number of local solutions. To accomplish this, we simply categorized our simulated data sets to have roughly the same univariate frequency distributions as our example data.

For each model fit to the simulated data, we generated 500 sets of random start values, using the same range of values as reported in Table 1. All models were again estimated with Mplus 3.01 using the accelerated EM algorithm. Because our findings for the two population models of the simulation were broadly similar, we discuss the results for each simultaneously.

Table 4
Results for Latent Class Growth Analysis (LCGA) Models and Growth Mixture Models (GMMs) Fitted With 500 Sets of Start Values to 200 Simulated Data Sets

Model	No. of estimated parameters	Start values that converged (%)[a]	No. of unique solutions[b]	Range of unique solutions[c]	Best solution (%)[d]	Data sets where modal solution has highest log-likelihood (%)[e]
LCGA						
3 classes	17	99.8	2.03	2–3	74.03	100.0
4 classes[f]	21	92.0	6.80	4–11	43.96	93.0
5 classes	25	86.6	16.37	9–26	13.26	21.0
GMM						
3 classes	23	94.4	5.85	3–10	5.37	0.0
4 classes[f]	27	92.3	9.72	4–15	77.02	100.0
5 classes	31	92.1	17.01	8–27	21.26	33.5

Note. The GMMs were fitted allowing random effects within classes.
[a] Percentage of start values leading to model convergence out of 500 randomized start values, averaged over 200 replications. [b] Number of different log-likelihood values found for model, averaged over 200 replications. [c] Minimum to maximum number of different log-likelihood values, over 200 replications. [d] Percentage of converged solutions that resulted in the maximum likelihood value, averaged over 200 replications. [e] Percentage of replications where highest log-likelihood solution was also the most common solution found for that model. [f] Fitted model is of the same form as the population generating model.

Specifying Too Many or Too Few Latent Classes

Number of unique solutions. Model complexity, as opposed to specification of the correct number of classes, was the key determinant of the number of local solutions obtained. Similar to the findings for our case study, with the simulated data we found that the number of unique solutions monotonically increased with the number of latent classes that were estimated. This occurred for both the LCGA and GMM specifications. Notably, even when estimating the proper four-class LCGA model, the number of unique solutions ranged from 4 to 11 across data sets and averaged more than 6 unique solutions per data set, as seen in Table 4. This effect for the GMM was even more dramatic: The number of unique solutions ranged from 4 to 15 when estimating the proper four-class model and averaged more than 9 unique solutions per data set, as seen in Table 4. This does not necessarily indicate that one can be less concerned about start values when conducting an LCGA, however, because in practice LCGA models will typically require more classes than GMMs to account for the same heterogeneity in change over time.[6]

Frequency of the optimal solution. Unlike the case study, in the simulation we obtained the solution with the highest log-likelihood fairly frequently when estimating the correct model. For instance, on average, the four-class LCGA solution with the highest log-likelihood was obtained in 44% of the random starts that converged. Additional probing of these results indicated minimal variability in this number across replications: In all replications the highest log-likelihood occurred in at least 39% of the random starts. We can also see in Table 4 that in 93% of the data sets for the LCGA model and in virtually all of the data sets for the GMM, the most frequently obtained solution was also the best solution. Nonetheless, we found evidence that a high frequency for the best solution was not always a reliable indicator of the correct model: In all 200 data sets, the percentage of random starts yielding the highest log-likelihood was always greater for the three-class LCGA model than it was for the proper four-class model. In contrast, whereas the solution with the highest log-likelihood for the four-class GMM was obtained, on average, from 77% of the random starts, the highest log-likelihood solution occurred relatively infrequently when estimating a GMM model with an *incorrect* number of classes. Further analyses revealed that for each data set, the solution for the model with the correct number of classes with the highest log-likelihood was also the solution with parameter estimates closest to those of the population-generating model. Thus, when estimating the correct model, we found no evidence in this simulation of a spurious solution with a better log-likelihood than that found for the actual proper solution.[7]

[6] Another interesting result of our simulation analysis was that a high rate of improper solutions (e.g., negative error variances, correlations outside the proper range) was obtained for the GMM even when the correct number of classes was specified (14% of the solutions). This suggests that improper estimates should not be used to reject a model and is consistent with the results of Chen, Bollen, Paxton, Curran, and Kirby (2001).

[7] We also found no evidence that the number of random starts that could be iterated to convergence was related to estimating the correct number of classes, as this number generally declined with increasing numbers of classes for both models.

Table 5
Results for Latent Class Growth Analysis (LCGA) Models and Growth Mixture Models (GMMs) Fitted With 500 Sets of Start Values to 200 Simulated Data Sets

Model	No. of estimated parameters	Start values that converged (%)[a]	No. of unique solutions[b]	Range of unique solutions[c]	Best solution (%)[d]	Data sets where modal solution has highest log-likelihood (%)[e]
LCGA						
Misspecified linear	17	33.6	4.24	2–9	14.23	0.0
Misspecified quadratic (linear time coding)	21	97.3	4.55	2–7	7.80	0.0
Properly specified quadratic[f]	21	92.0	6.80	4–11	43.96	93.0
GMM						
Misspecified linear	20	94.4	4.82	2–9	76.13	96.5
Misspecified quadratic (linear time coding)	27	97.4	9.17	3–16	80.01	100.0
Properly specified quadratic[f]	27	92.3	9.72	4–15	77.02	100.0

[a] Percentage of start values leading to model convergence out of 500 randomized start values, averaged over 200 replications. [b] Number of different log-likelihood values found for model, averaged over 200 replications. [c] Minimum to maximum number of different log-likelihood values, over 200 replications. [d] Percentage of converged solutions that resulted in the maximum likelihood value, averaged over 200 replications. [e] Percentage of replications where highest log-likelihood solution was also the most common solution found for that model. [f] The fitted model is of the same form as the population-generating model.

Model fit comparison. Similar to our case study, for the simulated LCGA and GMM data, the range of BIC values for the three- and five-class solutions overlapped with the four-class solutions. For the GMM, we found that on average across the 200 data sets, 42% of the four-class solutions were worse than the best three- or five-class solutions. This ranged from one data set in which only 5% of the four-class solutions were worse than the other models, to another data set in which this occurred for 75% of the four-class solutions. For the LCGA model, on average across the 200 data sets, about half of the four-class solutions were worse than the best three- or five-class solutions (ranging from 44% to 53% for any given data set). These results suggest that model selection could be adversely affected by inadvertently comparing local solutions.

Misspecifying the Within-Class Growth Model

We now briefly consider the effect of estimating the wrong within-class model when the number of classes is held constant at the correct number. For this comparison, we fit only GMMs to data simulated from a GMM model, but misspecified the growth process; we did the same for LCGA simulated data. Two misspecifications of these models were considered. In the first, the quadratic growth factor was omitted from the model. This not only represents a misspecification but also a decrease in model complexity, as fewer parameters must be estimated. As such, our second misspecification was to retain the form of the quadratic model, with the same number of parameters, but to use the incorrect time metric. Therefore, rather than specify the factor loadings according to the values in Equation 7, we used the more conventional linear metric of time:

$$\mathbf{\Lambda}_k = \begin{bmatrix} 1 & -.7 & .49 \\ 1 & -.6 & .36 \\ 1 & -.5 & .25 \\ 1 & -.1 & .01 \\ 1 & 0 & 0 \\ 1 & .5 & .25 \end{bmatrix}.$$

Note that the time values in the second column (for the linear effect) do not increment evenly because of the irregular spacing of the repeated measures.

Table 5 provides evidence that misspecifying the within-class model does not necessarily result in an increase in the number of local solutions. Indeed, in our simulation, the greatest number of solutions was obtained in both cases for the properly specified model. We also found evidence that in some instances, the optimal solution occurs just as frequently when estimating the incorrect within-class model as when estimating the correct population model. For both of the misspecifications with the GMM data (the linear model and the quadratic with time coded linearly), the optimal solution was nearly always the most frequent solution. However, caution should be exercised in generalizing this finding to other misspecifications: Given the complexity of these models, in practice researchers face the risk of misspecifying several different parts of the model at once. It thus remains an open question whether a greater degree of

misspecification could increase the number of local solutions for a model.

Misspecifying the Distribution of the Data

A final concern suggested by our case study is that perhaps the lack of alignment between the assumption of within-class normality and the ordinal nature of the data is the source of the many local solutions. To address this issue, we categorized our simulated data to resemble the empirical distributions in our case study (as detailed in the Appendix). Our results provide a preliminary indication that fitting a normal mixture to ordinal data (even with seven categories) can lead to optimization problems. Although the GMM still converged frequently, in only 9% of the data sets was the modal solution the best solution, and only 6.5% of the converged starts arrived at the best log-likelihood value (paralleling our findings with the example data), as seen in Table 6. It is also notable that the number of unique solutions increased dramatically when the GMM data were categorized. In contrast, when the population model was an LCGA model, the ordinal nature of the data had weaker effects: The model almost always converged, and the number of solutions did not increase relative to the continuous data. Nonetheless, the best model was also difficult to detect in the LCGA data: Only 20% of the converged starts arrived at the best log-likelihood value, and in only 20% of the data sets was the modal solution the best one. In addition to these implications for problems of estimation, violating the distributional assumptions of the model can create problems for model selection and interpretation, for instance, leading to the estimation of spurious classes (Bauer & Curran, 2003). In total, these considerations should motivate applied researchers to adopt GMM and/or LCGA models that are more appropriate for ordinal data whenever possible.

Summary

As in our case study, we found in the simulation that the percentage of starts that converged on a solution declined monotonically as a function of model complexity. Models with more classes, and those permitting random effects within classes, converged less frequently. The simulation also reinforced the finding from the case study that the number of unique solutions increases monotonically as more classes are added to the models. In addition, the simulation results replicated the case study finding of considerable overlap between models in the fit of local solutions (as judged by the BIC): This raises the possibility of selecting the wrong solution when failing to consider a wide range of start values.

It is notable that in our simulation, the optimal solution for the properly specified model occurred rather frequently. This suggests that knowing the frequency of the optimal solution may also help in identifying the true solution: If the optimal solution is found only rarely, this may suggest an error in the model specification. Given these results, we must more cautiously interpret the four-class solutions found for the GMM in the case study: Given that the optimal solution was found infrequently, this casts doubt on whether the four-class GMM was properly specified. Our simulation results suggest that the infrequency of this solution may be a consequence of incorrectly assuming normality for ordinal data, rather than a result of misspecification of the number of latent classes or within-class growth model.

We now consider the limitations of the present research and follow with a series of tentative recommendations for the estimation of GMMs in applied research.

Table 6
Results Obtained by Fitting Latent Class Growth Analysis (LCGA) Models and Growth Mixture Models (GMMs) With 500 Sets of Start Values to 200 Simulated Data Sets

Model	No. of estimated parameters	Start values that converged (%)[a]	No. of unique solutions[b]	Range of unique solutions[c]	Best solution (%)[d]	Data sets where modal solution has highest log-likelihood (%)[e]
LCGA						
Ordinal data	21	97.8	5.4	3–8	20.6	20.5
Continuous data	21	92.0	6.8	4–11	44.0	93.0
GMM						
Ordinal data	27	88.6	17.1	11–26	6.5	9.0
Continuous data	27	92.3	9.7	4–15	77.0	100.0

[a] Percentage of start values leading to model convergence out of 500 randomized start values, averaged over 200 replications. [b] Number of different log-likelihood values found for model, averaged over 200 replications. [c] Minimum to maximum number of different log-likelihood values, over 200 replications. [d] Percentage of converged solutions that resulted in the maximum likelihood value, averaged over 200 replications. [e] Percentage of replications where highest log-likelihood solution was also the most common solution found for that model.

Limitations and Future Directions

The approach we took here of a case study and a small Monte Carlo simulation has both positive and negative aspects. The core strength of the case study approach is that it permitted us to examine in depth analytical issues as they present themselves with real empirical data. A limitation of case studies is that their findings may not generalize to other models or data with other features. In general, however, the results we obtained in the case study were both predictable on the basis of analytical theory and were replicated in the small Monte Carlo simulation. Nonetheless, this simulation was only a preliminary foray intended to illustrate the sensitivity of these techniques to start values. We are hopeful that this article will generate interesting hypotheses to be tested in future Monte Carlo studies. A challenge for such studies is that they are computationally demanding. In general, the number of estimated models will equal the product of the number of conditions, the number of replications within condition, and the number of start values. Even in our small simulation, this totaled 1.2 million model runs.

Another valuable line of future research would be to develop more robust methods of model estimation. Two possibilities may be worth considering. First, some methods of estimation may be less likely to produce a local solution than maximum likelihood with the EM algorithm (or derivatives thereof). As pointed out by one reviewer, however, it is possible that approaches like simulated annealing or genetic algorithms would be so time consuming that using multiple start values would still be more efficient computationally. Second, one can start the optimization process at a better place. Our strategy for selecting start values could be described as naïve—randomly generating values from uniform distributions is probably not the most efficient way to select start values. Future work may help to increase the efficiency and the probability of locating the maximum likelihood solution for a given model by implementing more intelligent strategies for selecting start values (Biernacki, Celeux, & Govaert, 2003).

Recommendations

While GMMs are useful for modeling heterogeneity in change, many researchers may underestimate their complexity and the associated problems this can present to model estimation and selection. Although we are by no means the first to suggest that these models are susceptible to local solutions and that one should vary start values in estimating them, we believe that this is the first empirical evaluation of this issue to be presented to applied researchers. Without such studies, recommendations to vary start values are ambiguous. To some, this may indicate that if the same solution is obtained with three or four sets of start values, then it is likely the maximum likelihood solution (as opposed to a local optimum). Our results indicate that this will not always be the case and suggest that a much more extensive evaluation of the likelihood surface will often be necessary.

Given the consistency of our results with theoretical expectations and analytical work on related models, we feel that the present study offers a number of insights into how GMMs should be estimated in practice. We therefore conclude by offering three recommendations for applied researchers using these models.

Vary the Starting Values Extensively

While this advice has been given in the past, previous work has not been more specific. On one hand, our simulation showed that in an ideal instance where the model is properly specified and the data conform to the assumption of multivariate normality within classes, the best fitting model occurs at least 50% of the time. On the other hand, our case studies showed that in less ideal circumstances a GMM or LCGA may have to be estimated from many different sets of starting values to identify the solution with the highest log-likelihood. For instance, the best fitting model in our case study was obtained by only 3% of the random start values that led to model convergence, or about 1 in 30. Moreover, if we also consider random starts that failed to converge (81% for the four-class model), only about 1 in 167 of the random starts for the four-class model led to convergence on the solution with the highest log-likelihood. For complex models with many parameters, simply generating 10 sets of start values will be insufficient (the current default in Mplus 3 and Latent GOLD); instead, at least 50 to 100 sets of starting values will be needed. This becomes even more necessary with models containing more classes since they converge less frequently. Additionally, these start values must be sufficiently varied to fully probe the parameter space. While widely varying the start values may reduce the number of starts that lead to model convergence, this seems preferable to more narrowly varied start values that may fail to locate the solution with the highest log-likelihood.

Compare Various Solutions to Determine the Stability of the Model

Given that both our case study and Monte Carlo simulation found considerable overlap between models in the fit of local solutions (as judged by the BIC), failing to exhaustively explore the likelihood surface increases the chance that the selection of a final model (e.g., the best number of latent trajectory classes) will be misinformed. An important implication of this is that it is imperative to compare the substantive results of the key solutions obtained. In an instance where the top solutions all show substantively similar results, the researcher can be more confident in drawing substantive conclusions from the model. On the other hand, if solutions with similar log-likelihood values

diverge substantively, then this suggests that the results should be interpreted cautiously. Running models with more or fewer classes may help to determine the robustness of the latent classes. In some instances, there may simply not be enough information in the data set to choose with certainty between competing solutions. The sensitivity of the reported results to the choice of solutions should then be noted in the interpretation of the model.

Assess the Frequency of the Solution With the Highest Log-Likelihood as a Diagnostic

A tentative suggestion is to use the number of random starts converging to the solution with the highest log-likelihood as a diagnostic for the appropriateness of the model. While we found in our case study that the solution with the highest log-likelihood sometimes occurred quite infrequently, our small-scale simulation provided evidence that such rare solutions are unlikely if the correct model is specified. Thus, finding that the optimal solution occurs infrequently may be an indication that the model has been misspecified in some way. This implies a need for software development because it is not enough for the software to randomize the start values and then provide a single "best" solution. Instead, the researcher would need to allow a large number of start values to fully iterate to solutions and then view how frequently the optimal solution occurred. Even this is no guarantee, however, as our simulation showed that in some cases, the optimal solution occurred quite frequently even with misspecified models.

Software Considerations

Recent software developments make more feasible the strategy of varying the start values extensively, as both Latent GOLD and Mplus (beginning in Version 3) allow researchers to randomize the start values. Both of these programs use similar strategies: They randomly generate a certain number of sets of start values, iterate each for a particular number of iterations, and then choose the subset with the best likelihood values (at that point) to iterate to convergence. Neither software program fully documents the procedures used for randomizing start values, nor do these programs provide the user with full control over the randomization procedures (e.g., ranges for all parameter estimates).[8]

In Mplus, only the means of the latent factors are varied across classes. Our understanding is that random draws are taken from a uniform distribution extending ± 5 units from the start values initially input by the user (or zero by default). This range can be extended if desired, but whatever range is selected is applied to all estimated means. In the absence of manually input start values, the other model parameters are all given uniform start values across classes, regardless of whether the estimates are permitted to vary over classes in the specified model. The variances of the random growth factors are all given default start values of .05, whereas their covariances are started at zero. The residual variances are started at one half of the total variances of the repeated measures. Finally, the class proportions are all given equal start values, summing to 100%. Providing the user with greater control over the randomization process would be useful for two reasons. First, without randomizing the other parameter estimates of the model, the default procedure probes only a limited area of the parameter space. Second, using a uniform distribution of constant width to generate the start values for all of the mean estimates is reasonable only if the means are scaled commensurately. This condition does not hold for GMMs, where a range of ± 5 could be small for intercepts, large for linear slopes, and enormous for quadratic parameters.

The algorithm used to generate random start values in Latent GOLD is more complex (J. Magidson, personal communication, June 18, 2005). Random start values are generated for all of the parameters in the model. Start values are generated by first computing a function of two random values, a random number from a uniform distribution between $-.5$ and .5, and an *extremeness factor,* which is an integer from 1 to 5 (with equal probability). A different uniform value is drawn for each start value in the model, but only one extremeness factor is drawn for each set of start values. The precise function used to scale the random uniform by the extremeness factor depends on the type of parameter (e.g., mean, variance, class probability). The resulting value is then combined with information about the means and/or variances of the observed variables to produce the random start values, where again the combination depends on the type of parameter. The end result is that the dispersion of the start values is tuned to the specific type of parameter in the model. Although this is a clear advantage of the Latent GOLD algorithm, a disadvantage is that the user has no apparent control over the range of start values generated for the parameters of the model.

In contrast to Mplus and Latent GOLD, the PROC TRAJ macro currently requires the manual input of all start values (Jones, Nagin, & Roeder, 2000). It is a clear disadvantage of this program that no automated randomization of start values is available, yet the fact that PROC TRAJ is implemented within the SAS data system offers the user other opportunities to address local solutions. Namely, the SAS macro language can be used to randomize the start values, fit the model, and compile the results across different random starts in much the same way that we did for the analyses reported here.

[8] To determine how Mplus 3.01 generates starting values, we tricked the program by estimating a number of models with zero iterations. This provided information on the values the program uses.

Thus, while the incorporation of start value randomization procedures into current software is a welcome addition, the current defaults are probably insufficient, both in terms of the number of start values that are generated (10 in Mplus and Latent GOLD, zero in PROC TRAJ) and in terms of the extent of control provided to the user over the range of values. Indeed, for our simulated data, we estimated four-class GMM models using the defaults of Mplus 3.01 and found that the correct solution was obtained in only 23% of the 200 data sets. Although the Mplus manual suggests increasing the number of random starts for "a more thorough investigation of multiple solutions" (L. K. Muthén & Muthén, 2004, p. 379), it is hard to conceive of an instance in which it would be wise for a researcher to actually use these default values. Therefore, it is imperative that the researcher override these default values to specify a much greater range of start values. Along these lines, an important advantage of Latent GOLD is that one can store new default values for subsequent model runs.

Conclusion

In closing, GMMs offer an exciting new way to evaluate theories concerned with population subgroups showing different patterns of change over time. Yet these models also bring new complexities that applied researchers may be unfamiliar with. Given the preponderance of local solutions for these models, one simply cannot input the data, specify a model, and immediately interpret the results of a single model run. The magnitude of this issue is typically not communicated in published empirical articles presenting GMMs and LCGAs, perhaps because such articles lack the space to detail all of the ancillary analyses that were performed. By thoroughly evaluating the presence and impact of local solutions in an empirical data set and a small-scale simulation, the present article demonstrated that start values must be taken seriously when estimating GMMs, whether or not random effects are included. This will often require overriding the defaults of commonly used software programs.

References

Aldwin, C. M., Spiro, A. I., Levenson, M. R., & Cupertino, A. P. (2001). Longitudinal findings from the Normative Aging Study: III. Personality, individual health trajectories, and mortality. *Psychology and Aging, 16,* 450–465.

Arminger, G., & Stein, P. (1999). Finite mixture of covariance structure models with regressors: Loglikehood function, distance estimation, fit indices, and a complex example. *Sociological Methods & Research, 26,* 148–182.

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

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

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

Biernacki, C. (2003). *Testing for a global maximum of the likelihood.* Unpublished manuscript.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis, 41,* 561–575.

Biernacki, C., & Chretien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM. *Statistics & Probability Letters, 61,* 373–382.

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective.* New York: Wiley.

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

Day, N. E. (1969). Estimating the components of a mixture of two normal distributions. *Biometrika, 56,* 463–474.

Dolan, C. V., Jansen, B. R. J., & van der Maas, H. L. J. (2004). Constrained and unconstrained multivariate normal finite mixture modeling of Piagetian data. *Multivariate Behavioral Research, 39*(1), 69–98.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997a). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*(1), 39–59.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997b). STEMM: A general finite mixture structural equation model. *Journal of Classification, 14*(1), 23–50.

Jones, B. L., Nagin, D. S., & Roeder, K. (2000). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29,* 374–393.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–906.

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

Magidson, J., & Vermunt, J. K. (2004). Latent class analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 175–198). Thousand Oaks, CA: Sage.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *The 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*(1), 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., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Meredith, W., & Tisak, J. (1984, July). *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*(1), 107–122.

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

Molina, C. G. (2000). Assessing the number of components in finite Gaussian mixtures by generalised Fisher ratio, normalised entropy criterion and functional merging. *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, 26*(1–2), 95–103.

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

Muthén, B. O., & 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. O., & Muthén, L. K. (2001). *Mplus: Statistical analysis with latent variables user's guide* (2nd ed.). Los Angeles: Stat Model.

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

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide: Third Edition.* Los Angeles: Authors.

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

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.

Schulenberg, J., O'Malley, P., Bachman, J., Wadsworth, K., & Johnston, L. (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.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis.* New York: Oxford.

Solka, J. L., Wegman, E. J., Priebe, C. E., Poston, W. L., & Rogers, G. W. (1998). Mixture structure analysis using the Akaike information criterion and the bootstrap. *Statistics and Computing, 8*(3), 177–188.

Steinley, D. (2003). Local optima in K-means clustering: What you don't know may hurt you. *Psychological Methods, 8,* 294–304.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*(433), 217–221.

Vermunt, J. K., & Magidson, J. (2005). *Latent Gold Choice 4.0 user's manual.* Belmont, MA: Statistical Innovations.

Vermunt, J. K., & van Dijk, L. A. (2001, December). A Nonparametric random coefficient approach: The latent class regression model. *Multilevel Modelling Newsletter, 13*(2), 6–13.

Appendix

Data Generation Methodology for the Monte Carlo Study

For our simulation study, all data generation was performed in SAS 8.02 using the PROC IML matrix algebra programming language. We used the parameter estimates originally presented by B. O. Muthén and Muthén (2000) as the population parameters for the generating model. This included the parameter estimates for each class, as well as the class proportions. The sample size was set to the number of cases used in the original analysis, or $N = 924$. A total of 200 replications were simulated for each of the two population models considered in the simulation, namely the GMM with random effects and the LCGA. Here we describe the method used to generate the data for a given replication.

Data generation was accomplished in two stages. First, to determine class membership, we sampled on the basis of the class probabilities. The class probabilities were $\pi_1 = .734$, $\pi_2 = .070$, $\pi_3 = .050$, and $\pi_4 = .146$. We used the cumulative probabilities (i.e., .734, .804, .853, and 1.000) to assign threshold points along a uniform distribution ranging from 0 to 1 and then sampled from this uniform distribution to assign cases to classes (using the RANUNI function in SAS). A draw less than the first threshold would result in assignment to Class 1, a draw between the first and second thresholds would result in assignment to Class 2, and so on. For example, a draw of .82 would place a case in Class 3. The obtained number of cases for each class is designated as n_k, where $N = \Sigma n_k$.

In the second stage of data generation, we simulated the observed vector of repeated measures $\mathbf{y}_i$, for each case given its class membership. Specifically, for each class k, the $p = 6$ repeated measures were sampled from a multivariate normal distribution given by

$$\mathbf{y}_i | k \sim MVN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ were parameterized as

$$\boldsymbol{\mu}_k = \boldsymbol{\Lambda}\boldsymbol{\alpha}_k,$$

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

(here we omit the θ from Equation 6 in the text for compactness). Notice that only $\boldsymbol{\alpha}_k$, the mean vector of the growth parameters (intercept, linear component, quadratic component), varied across classes. This vector was differentially defined for the four classes as

$$\boldsymbol{\alpha}_1' = [4.56 \quad -.869 \quad -.562]$$

$$\boldsymbol{\alpha}_2' = [.353 \quad -.003 \quad -.017]$$

$$\boldsymbol{\alpha}_3' = [1.63 \quad -.682 \quad -.006]$$

$$\boldsymbol{\alpha}_4' = [2.45 \quad -.042 \quad -.228].$$

The model-implied means of the repeated measures defined by these values and $\boldsymbol{\Lambda}$ (as given previously by Equation 7 in the text) are plotted in Figure 4B.

To obtain the covariance matrix of the repeated measures, the covariance matrix of the random effects was set to

$$\boldsymbol{\Psi} = \begin{bmatrix} .739 & -.271 & -.166 \\ -.271 & .130 & .073 \\ -.166 & .073 & .043 \end{bmatrix}$$

for the random effects GMMs and was set to $\boldsymbol{\Psi} = \mathbf{0}$ for the LCGA models. Finally, the residual covariance matrix $\boldsymbol{\Theta}$ taken from the analysis of B. O. Muthén and Muthén (2000) was as follows:

$$\boldsymbol{\Theta} = DIAG[0.062 \quad 1.253 \quad 1.357 \quad 1.102 \quad 1.123 \quad 0.988].$$

To generate data with the given within-class moment structure, we first independently sampled p random normal deviates for each of the n_k cases assigned to a given class from a standard normal distribution (using the RANNOR function in SAS) to produce an $n_k \times p$ matrix $\mathbf{Z}_k$. To impose the specified covariance structure, we multiplied $\mathbf{Z}_k$ by the Cholesky root of $\boldsymbol{\Sigma}$ and then added the constant $\boldsymbol{\mu}_k$ such that

$$\mathbf{Y}_k = \mathbf{Z}_k\mathbf{L} + \mathbf{1}\boldsymbol{\mu}_k',$$

where $\mathbf{L}$ is the Cholesky root of $\boldsymbol{\Sigma}$ and $\mathbf{1}$ is an $n_k \times 1$ vector of ones. This process was repeated for each of the k classes, and then the total data set was assembled by stacking the data matrices for the four classes, $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$, and $\mathbf{Y}_4$, into a single $N \times p$ matrix $\mathbf{Y}$. This aggregated data matrix was saved to an external file for subsequent analysis in Mplus. Different random number seeds were used for the random draws for each of the 200 replications in both conditions to produce sampling variability in the class proportions and the within-class means and covariances of the repeated measures.

For some aspects of the simulation, we subsequently categorized the data. To do so, we used the following thresholds for all six time points for both of our simulated models: $\tau_1 = 1.323$, $\tau_2 = 1.812$, $\tau_3 = 2.567$, $\tau_4 = 3.248$, $\tau_5 = 3.715$, and $\tau_6 = 4.199$. If the continuous observed values generated in the preceding step were less than the first threshold, then they were set to 0, if they were between the first and second threshold, they were set to 1, and so on. These thresholds produced univariate frequency distributions that closely matched the observed frequency distributions in the actual National Longitudinal Survey of Youth data.

Received March 17, 2004
Revision received June 29, 2005
Accepted October 24, 2005 ■