MULTIVARIATE BEHAVIORAL RESEARCH, *42*(4), 757–786

2004 Cattell Award Address

Observations on the Use of Growth Mixture Models in Psychological Research

Daniel J. Bauer
The University of North Carolina at Chapel Hill

Psychologists are applying growth mixture models at an increasing rate. This article argues that most of these applications are unlikely to reproduce the underlying taxonic structure of the population. At a more fundamental level, in many cases there is probably no taxonic structure to be found. Latent growth classes then categorically approximate the true continuum of individual differences in change. This approximation, although in some cases potentially useful, can also be problematic. The utility of growth mixture models for psychological science thus remains in doubt. Some ways in which these models might be more profitably used are suggested.

Growth mixture models (GMMs) are designed to separate a general population of individuals into subgroups characterized by qualitatively distinct patterns of change over time. In this article, I offer a few observations on the application of these models in psychological science. Like many, I was initially excited about the potential of GMMs. After several years of evaluating these models and reviewing applications, however, I am now skeptical that they will meaningfully advance our understanding of psychosocial development. In what follows, I outline key methodological and theoretical concerns that I have with current applications of GMMs.

Correspondence concerning this article should be addressed to Daniel Bauer, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270. E-mail: dbauer@email.unc.edu

SOME CONTEXT

Every now and then a new statistical model is developed that captivates the field. In the area of longitudinal data analysis, models for mean change, such as ANOVA and MANOVA, quickly went out of favor in the 1990s as individual trajectory models, like the latent curve model (LCM) and multilevel growth model, were shown to have important advantages. Now, GMMs seem to be gaining favor for their ability to parse unobserved heterogeneity in change over time. Go to a conference on developmental psychopathology and you'll see any number of longitudinal data sets analyzed by GMMs.

The idea of clustering trajectories is not new to psychology. McCall, Appelbaum, and Hogarty (1973) used a sophisticated clustering procedure to differentiate patterns of IQ development 2 decades before the debut of GMMs. But cluster analysis never gained much momentum, perhaps because many cluster analyses are performed via heuristic algorithms (e.g., minimizing squared Euclidean distances). These algorithms will produce clusters from any data, even if the data is generated randomly. Given this, an analyst's claim that "cluster analysis of the data revealed four clusters" is hardly credible. The analyst, not the algorithm, selected four clusters because four clusters seemed to be a good number for the data. Of course, had the analyst used a different clustering rule, entirely different clusters might have emerged. To many scientists, cluster analysis is thus tainted by an unavoidable subjectivity. Indeed, one prominent psychologist, sympathetic to the general idea of clustering, remarked to me that he could not help but feel that cluster analysis was a little like reading tea leaves.

Perhaps one of the greatest attractions of GMMs is that they appear to provide a more principled, objective approach for identifying trajectory groups (Connell & Frye, 2006). In contrast to heuristic algorithms, the clusters are specified as part of a formal statistical model with parameters estimated using conventional methods like maximum likelihood. To their credit, Daniel Nagin and his colleagues were the first social scientists to suggest this approach for clustering trajectories (Land & Nagin, 1996; Nagin & Land, 1993; Nagin & Tremblay, 1999). In part, their inspiration was Moffitt's (1993) influential theory on the development of antisocial and criminal behavior. Moffitt posited a developmental taxonomy consisting of two etiologically distinct groups: one group whose antisocial behavior onsets in adolescence and then desists in early adulthood, and another group that exhibits childhood onset and persistent antisocial behavior throughout adulthood. Quite laudably, Nagin and Land sought to develop a statistical model that could be used to more adequately evaluate this theory and others like it.

In later publications, Nagin (1999) and B. Muthén & Shedden (1999) presented GMMs for general consumption and offered user-friendly software for fitting the models. These were positive contributions, but there were also unintended consequences. Focused on promoting the possibilities of the models, and their new accessibility to applied researchers, neither Nagin nor B. Muthén dwelled much on the assumptions or potential limitations of GMMs. The rapid pace of software development also meant that GMMs were delivered to applied researchers well before independent methodologists could critically examine the robustness of the models in peer-reviewed publications. Perhaps not unreasonably, users trusted that GMMs could divine both the number of groups and the shapes of the group trajectories, even when there were no specific taxonic hypotheses to motivate the analysis.

I, too, was initially enthusiastic about GMMs. When a fellow graduate student and I attended the conference *New Methods for the Analysis of Change* in 1998, it was Bengt Muthén's presentation on GMMs that we found most engaging. Here was a method that could be used to tease apart population heterogeneity in change over time. No longer would we be limited to single trajectory models like latent curve analysis—now we could allow for multiple trajectories. And given the choice between a single trajectory and multiple trajectories, who would not opt for the latter? Surely, no population is completely homogeneous. In retrospect, it's clear that, in my naiveté, I was imposing a false dichotomy on the models. LCMs do not imply one trajectory, they imply many trajectories. But at the time I was simply too caught up in the excitement of these new and improved growth models to think through their real benefits or possible costs.

I was so excited, in fact, that I set about writing a post-doctoral fellowship grant with Kenneth Bollen and Patrick Curran focused on studying GMMs in more detail. The aims of the proposal were to study the analytics of the GMM, to consider the model's performance with simulated data, and then to apply the model with real data. Soon after I began the fellowship, I attended a symposium on advances in modeling individual development at the 2001 meeting of the Society for Research in Child Development. One of several excellent presentations was Bengt Muthén's talk on GMMs. At the end of his talk, Dr. Muthén made an offhand remark that there was a great deal of "low hanging fruit" ripe for the plucking by graduate students and post-docs looking for dissertation or fellowship topics related to GMMs. The discussant, Mark Appelbaum, then quipped that we had better watch out—low hanging fruit is often rotten.

Those words were apropos. My initial exuberance for GMMs had been slowly turning to pessimism. To be sure, GMMs could recover subgroups characterized by different patterns of change over time under just the right conditions. But they could also provide evidence of illusory subgroups when assumptions of the model were incorrect. Further, it seemed to me those assumptions would be incorrect most of the time in real-data applications. No wonder, then, that I stopped short of fulfilling the last aim of my fellowship, to conduct an empirical analysis using a GMM.

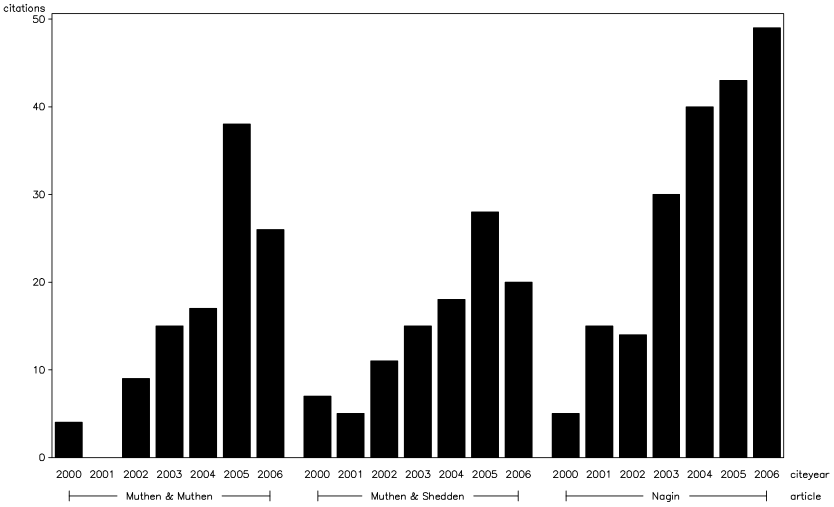

FIGURE 1 Number of citations by year for key early papers on growth mixture modeling in psychological research by B. O. Muthén & Muthén (2000), B. Muthén & Shedden (1999), and Nagin (1999).

In the years since, I have seen the number of applications of GMMs continue to rise in the literature. To put a bit of data behind this observation, I performed a citation search for three early papers introducing GMMs to applied researchers, namely, Nagin (1999), B. Muthén and Shedden (1999), and B. O. Muthén and Muthén (2000). The total citation counts for the three papers were, respectively, 200, 105, and 109 citations, and the overall trend shown in Figure 1 points to increasing use of GMMs (Web of Science, Science Citation Index, retrieved 4/8/07).[1] The cautions voiced by myself and others concerning the application and interpretation of these models (e.g., Bauer & Curran, 2003a, 2003b, 2004; Eggleston, Laub, & Sampson, 2004; Hoeksma & Kelderman, 2006; Raudenbush, 2005; Sampson & Laub, 2005; Sampson, Laub, & Eggleston, 2004) seem to have gone largely unheard. The purpose of this article is to try to convey, more convincingly, my concerns about the use of GMMs in psychological research.

[1] In comparison, two foundational papers on latent curve modeling, Meredith & Tisak (1990) and McArdle & Epstein (1987), were cited 171 and 119 times, respectively, between 1999 and 2006. A key paper describing how longitudinal data could be analyzed by way of hierarchical linear models (mutlilevel models) published by Bryk & Raudenbush (1987) was cited 283 times over the same period.

METHODOLOGICAL CONCERNS

The ideal application of a model should meet three criteria. First, the core assumptions of the model should be met, approximately. Alternatively, the model should be known to provide robust inference despite the violation of some model assumptions. Second, the model should provide a reasonable representation of reality. That is, the model should resemble, as much as possible, the true underlying structure of the data. Third, the results obtained from the model should be scientifically meaningful (e.g., provide tests of specific, disconfirmable hypotheses). In this section, I focus on the first of these criteria by describing the core assumptions of the GMM and then considering whether these assumptions are met in practice. When assumptions are not met, the sensitivity of the results, particularly the number of estimated latent classes, will be of key concern. For contrast, the LCM is presented first and this is then extended to a GMM.

The Latent Curve Model

The LCM (McArdle, 1988; McArdle & Epstein, 1987; Meredith & Tisak, 1984, 1990) may 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 repeated measures for person i, $\boldsymbol{\eta}_i$ is a $q \times 1$ vector of latent trajectory parameters (e.g., intercept and slope), $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, and $\boldsymbol{\varepsilon}_i$ is a $p \times 1$ vector of residuals (usually assumed to be independent over time). In most applications, the values of the factor loadings are fixed to particular values to indicate that the individual trajectories follow a specific function (Bollen & Curran, 2006; Browne, 1993). Often (but not always), the $p \times p$ covariance matrix for $\boldsymbol{\varepsilon}_i$, designated $\boldsymbol{\Theta}$, is also assumed to be diagonal, implying that the correlations among the repeated measures are entirely accounted for by the underlying growth factors.

To complete the specification of the LCM we must also write a model for the latent variables. In the simplest case, the latent variables can be expressed in mean deviation form as

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\zeta}_i, \tag{2}$$

where $\boldsymbol{\alpha}$ is a $q \times 1$ vector holding the means of the individual trajectory parameters and $\boldsymbol{\zeta}_i$ is a $q \times 1$ vector of residuals, or random effects, reflecting individual differences from these means (assumed to be uncorrelated with $\boldsymbol{\varepsilon}_i$). Note that the inclusion of the random effects in the model implies that there are infinitely many possible individual trajectories in the population and that each individual follows a unique trajectory.

Substituting Equation 2 into Equation 1, we obtain the following equation:

$$\mathbf{y}_i = \boldsymbol{\Lambda\alpha} + \boldsymbol{\Lambda\zeta}_i + \boldsymbol{\varepsilon}_i, \tag{3}$$

which is sometimes referred to as an unconditional LCM. If we make the assumptions that $\boldsymbol{\zeta}_i$ and $\boldsymbol{\varepsilon}_i$ are normally distributed, then this implies that the marginal distribution of the repeated measures is

$$f(\mathbf{y}_i) = \phi\left(\boldsymbol{\Lambda\alpha}, \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}\right), \tag{4}$$

where ϕ is a p-dimensional normal probability density function. Maximum likelihood (ML) may then be used to estimate the parameters of the model. Note that the ML estimates are robust to certain forms of non-normality (Browne, 1984; Browne & Shapiro, 1988; Satorra, 1990; Satorra & Bentler, 1990).[2] Other estimators for these models also exist that forgo the normality assumption (Browne, 1984; Yuan, Bentler, & Chan, 2004).

If predictors of change are present, then they may be incorporated by augmenting the latent variable model as follows:

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i, \tag{5}$$

where $\mathbf{x}_i$ is an $r \times 1$ vector of exogenous predictors and $\boldsymbol{\Gamma}$ is a $p \times r$ matrix of regression coefficients relating the predictors to the latent variables. This model is often referred to as the conditional LCM.

Similar to before, if we substitute Equation 5 into Equation 1, we obtain the following model for the observed repeated measures:

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\Lambda}\left(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i\right) + \boldsymbol{\varepsilon}_i \qquad (6) \\ &= \boldsymbol{\Lambda\alpha} + \boldsymbol{\Lambda\Gamma}\mathbf{x}_i + \boldsymbol{\Lambda\zeta}_i + \boldsymbol{\varepsilon}_i. \end{aligned}$$

Like most structural equation models, LCMs are typically estimated from the joint distribution of $\mathbf{y}_i$ and $\mathbf{x}_i$. However, to make the extension to the GMM, we instead consider the conditional distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$. For a typical structural equation model, Jöreskog (1973) showed that both the joint and conditional likelihoods yield equivalent estimates when $\mathbf{x}_i$ is exogenous. However, Arminger, Stein, and Wittenberg (1999) demonstrated that the conditional formulation is preferable for mixture models.[3]

[2] The standard errors of the ML estimates can be biased when the data are not multivariate normal, but robust standard errors can be computed via the method of Satorra & Bentler (1994).

[3] In contrast to the LCM, this conditional formulation is standard for multilevel growth models.

If we make the assumption that both $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\zeta}_i$ are normally distributed in Equation 6, then $\mathbf{y}_i$ has the probability density function

$$f(\mathbf{y}_i|\mathbf{x}_i) = \phi\left(\boldsymbol{\Lambda\alpha} + \boldsymbol{\Lambda\Gamma}\mathbf{x}_i, \boldsymbol{\Lambda\Psi\Lambda}' + \boldsymbol{\Theta}\right). \tag{7}$$

As before, this expression leads naturally to a maximum likelihood estimator for the model. Again, the ML estimates are robust to certain forms of non-normality.

Note that there are many possible ways to extend the simple LCMs presented here. One can, for instance, include time-varying predictors, latent predictors, or latent outcomes in the model, among other possibilities (see Bollen & Curran, 2006). The key extension that is the focus of this article, however, is to allow for heterogeneity in change over time across latent subgroups in the population.

The Growth Mixture Model

The key notion of the GMM is that the population is composed of K latent classes, each characterized by its own LCM. Each class thus has its own mean trajectory, and usually it is the mean trajectory that is used to name the class (e.g., "high-chronic" or "low-increasing"). Systematic individual differences from the mean trajectory within classes may be allowed, as in B. Muthén & Shedden (1999), or not, as in Nagin (1999). Here I present the model in a general way, later noting common restrictions used in practice.

For the unconditional GMM, we elaborate Equation 4 to account for the multiple latent classes mixed together in the population as follows:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k \left(\boldsymbol{\Lambda}_k\boldsymbol{\alpha}_k, \boldsymbol{\Lambda}_k\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k\right). \tag{8}$$

The distribution of the repeated measures is thus a finite mixture of normal distributions with means and covariances structured similarly to an LCM (relying on the assumption that $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\zeta}_i$ are normally distributed within-class, as in Equation 4). All parameters have the same interpretation as in the LCM. Accordingly, differences in $\boldsymbol{\Lambda}_k$ correspond to class differences in trajectory form, differences in $\boldsymbol{\alpha}_k$ capture differences in the mean trajectories of the classes, and differences in $\boldsymbol{\Psi}_k$ and $\boldsymbol{\Theta}_k$ capture differences in the dispersion of the individual trajectories and time-specific residuals within classes, respectively. The new parameter, π_k, represents the probability that a participant belongs to class k and is also interpretable as the proportion of individuals within the population from class k.

To be general, Equation 8 includes the subscript k after each parameter matrix to indicate that all of the parameters can vary over latent classes. Constraints are often imposed, however, to simplify the model. Commonly, the form of the individual (and mean) trajectories is assumed to be equal such that $\mathbf{\Lambda}_k = \mathbf{\Lambda}$. The dispersion matrices may also be held equal over classes, that is, $\mathbf{\Psi}_k = \mathbf{\Psi}$ and $\mathbf{\Theta}_k = \mathbf{\Theta}$. Seldom motivated by theory, these constraints are most often imposed for statistical expedience (e.g., to guarantee a global maximizer; see Hipp & Bauer, 2006). Alternatively, assuming that there is no within-class variability in the individual trajectories (i.e., $\mathbf{\Psi}_k = \mathbf{0}$) produces a model similar to Nagin's (1999).

At this point it is worth contrasting three alternative ways of representing individual differences in change over time with these models. If there is but one class, the GMM reduces to the LCM and all individual differences in change are within-class differences captured by the dispersion matrix $\mathbf{\Psi}$. Alternatively, if we fit a Nagin-type model (e.g., where $\mathbf{\Psi}_k = \mathbf{0}$ and $K > 1$), then differences among the individual trajectories must be captured entirely by between-class differences, represented in the vectors $\boldsymbol{\alpha}_k$. Finally, in the general GMM (e.g., where $\mathbf{\Psi}_k \neq \mathbf{0}$ and $K > 1$), individual differences in change over time are decomposed into a between-class component (differences in $\boldsymbol{\alpha}_k$) and a within-class component (the dispersion matrix $\mathbf{\Psi}_k$).

Because in general, individual differences in change over time are separated into two parts in the GMM, we must consider both parts when adding predictors to the model. Prediction may be either of class membership (the between-class component) or of relative standing within class (the within-class component), or both. Therefore, we elaborate Equation 8 to be a conditional GMM of the form

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i)\phi_k\left(\mathbf{\Lambda}_k\boldsymbol{\alpha}_k + \mathbf{\Lambda}_k\mathbf{\Gamma}_k\mathbf{x}_i, \mathbf{\Lambda}_k\mathbf{\Psi}_k\mathbf{\Lambda}'_k + \mathbf{\Theta}_k\right), \tag{9}$$

where

$$\pi_{ik}(\mathbf{x}_i) = \frac{\exp\left(\alpha_{c_k} + \boldsymbol{\gamma}'_{c_k}\mathbf{x}_i\right)}{\sum_{k=1}^{K}\exp\left(\alpha_{c_k} + \boldsymbol{\gamma}'_{c_k}\mathbf{x}_i\right)}. \tag{10}$$

The two key changes in the model, relative to Equation 8, are to make the class probabilities a multinomial function of $\mathbf{x}_i$ in Equation 10 (between-class prediction)[4] and to make the class means of the trajectory parameters conditional on the covariates through the linear function $\mathbf{\Lambda}_k\boldsymbol{\alpha}_k + \mathbf{\Lambda}_k\mathbf{\Gamma}_k\mathbf{x}_i$ (within-class

[4] For identification, one of the classes is declared as the reference class and the intercept (α_{c_k}) and slopes ($\boldsymbol{\gamma}_{c_k}$) of the multinomial regression for this class are set to zero. For example, if the last class were the reference class, then $\alpha_{c_K} \equiv 0$ and $\boldsymbol{\gamma}_{c_K} \equiv \mathbf{0}$. For interpretation, the slope estimates are typically exponentiated to provide odds ratios indicating the increased odds of being in class k relative to the reference class for each unit change in x.

prediction). Just as with the unconditional GMM, a variety of constraints are possible for this model.

Assumptions Met or Mangled?

The formulation and fitting of GMMs requires a number of assumptions. Here I enumerate some of the key assumptions of the model, consider how often these assumptions are met in practice, and survey results on the robustness of the GMM when specific assumptions are violated. I am particularly concerned with recovery of the latent class structure, which is often used to develop and evaluate taxonomic theories, identify problematic subgroups, and motivate targeted interventions. My conclusions about whether or not assumptions are typically met in practice are based on a survey of 21 articles published in 2005 that cited B. O. Muthén & Muthén (2000). Most of these articles focused on substance use (alcohol, cigarettes, marijuana, etc.), aggression/conduct problems, or depression, and most reported three or four latent trajectory classes.

Assumption 1: Within-class conditional normality. Most GMMs assume that the repeated measures are normally distributed within classes (conditional on any exogenous predictors), although it is possible to modify this assumption for certain types of outcomes (e.g., binary, ordinal, or count outcomes). If there is more than one class, then the assumption of within-class normality implies that the marginal distribution of the repeated measures (pooled over classes) must be non-normal. Because the latent classes are unobserved, it is this non-normality that is used to infer the presence of the latent classes and recover their characteristics. But non-normality can arise through other mechanisms that do not imply the existence of underlying population subgroups. In particular, in virtually all of the 21 applications I reviewed, limitations of measurement precluded the possibility of obtaining normally distributed repeated measures, irrespective of whether or not latent subgroups were mixed together in the population.

In several applications, the outcome was measured as a sum of several proportions (reflecting peer nominations). For proportions, normality will only be approximately attained if the denominators are large and there is a mid-range probability of endorsement, neither of which was the case in these applications. Linear composites were also common. Such composites, however, are often skewed and have pronounced floor effects for domains like substance use and aggression. In addition, log-transformed counts made several appearances. Here, the log transformation might aid skew but would be unlikely to correct for the "piling up" of zeros that is typically seen with this form of measurement. Finally, the outcome variable was in some cases ordinal. Even with many categories, however, interval-level spacing is far from assured. Thus, mixture or not, the assumption of within-class normality was in most applications an impossibility. Despite this, in only one case did the authors select a GMM explicitly formulated for non-normal outcomes (due to the binary nature of the dependent variable).

As shown by Bauer & Curran (2003a), if the marginal distribution of the repeated measures is non-normal, a GMM with multiple classes will almost always fit better than a single-group LCM, regardless of the source of non-normality. The problem arises because a mixture of normal distributions can approximate a variety of non-normal shapes, even in the absence of true population subgroups. This is shown in Figure 2 for a sequence of five repeated measures. The figure contrasts a non-parametric kernel density estimate with the density curve implied by a 2-class GMM over five occasions of measurement (for details, see Bauer & Curran, 2003b). Although the mixture closely approximates the kernel estimator, the latent class structure does not reflect the actual organization of individuals within the population: In this case, the data were generated from a unitary (but non-normal) distribution, not a mixture of normal distributions. Of course, distortions of the latent class structure would also arise if the data had been generated from a mixture of non-normal distributions and then fit by a mixture of normals (Hoeksma & Kelderman, 2006; Tofighi & Enders, 2007).

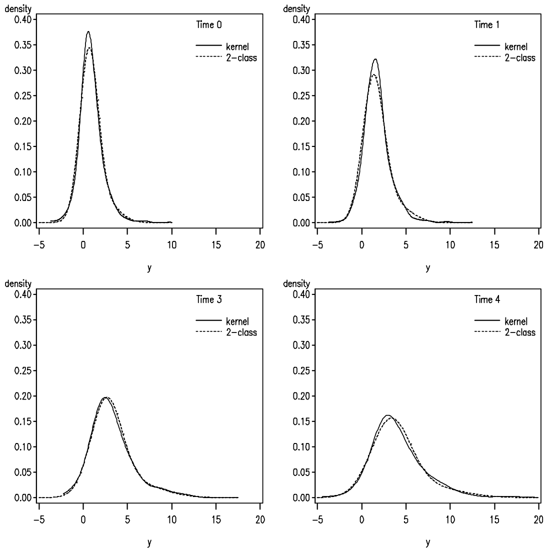

FIGURE 2 Kernel density estimates for five non-normal repeated measures simulated to have skew 1.5 and kurtosis 6 superimposed with the implied density obtained from a two-class growth mixture model with normal component distributions.

Given the preponderance of non-normal data in psychological research (Micceri, 1989), this problem is particularly vexing.

The conundrum is this: Our measurement instruments are typically incapable of producing observations that meet the conditional normality assumption, yet even mild violations of this assumption will typically result in the estimation of too many latent classes (Bauer & Curran, 2003a; Tofighi & Enders, 2007). This clearly compromises the usefulness of the GMM for evaluating taxonomic theories, as hypotheses of population heterogeneity will almost never be disconfirmed even when they are wrong. More broadly, the chance that the estimated latent classes will accurately map onto distinct population subgroups seems rather remote, even if such subgroups truly exist. The odds of identifying the correct latent class structure would improve if better measures could be developed or if GMMs for count or ordinal outcomes were implemented when appropriate.

Assumption 2: Properly specified mean and covariance structure. Another assumption of the GMM is that, within classes, the growth model accurately reproduces the means, variances, and covariances of the repeated measures. This assumption is rarely assessed in practice. In a few of the applications I reviewed, reasonable model fit was established for the standard LCM before proceeding to the GMM. Others, however, justified use of a GMM from the poor fit of the LCM. Although this conjecture might seem reasonable, the poor fit of the LCM could also be indicative of some other misspecification that the latent classes simply serve to cover up.

This issue can be clarified by considering the covariance structure implied by the model. Drawing on Bauer & Curran (2004), the covariance matrix of the aggregate repeated measures data implied by an unconditional GMM can be expressed as

$$\begin{aligned}\boldsymbol{\Sigma} &= \sum_{k=1}^{K}\sum_{j=k+1}^{K} \pi_k \pi_j \left(\boldsymbol{\Lambda}_k \boldsymbol{\alpha}_k - \boldsymbol{\Lambda}_j \boldsymbol{\alpha}_j\right)\left(\boldsymbol{\Lambda}_k \boldsymbol{\alpha}_k - \boldsymbol{\Lambda}_j \boldsymbol{\alpha}_j\right)' \\ &\quad + \sum_{k=1}^{K} \pi_k \left(\boldsymbol{\Lambda}_k \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k\right).\end{aligned} \tag{11}$$

Notice that this represents an additive decomposition of the overall covariance matrix into a portion due to the class mean differences, the first term, and a portion due to within-class covariance, the second term. As a result, misspecification of the within-class covariance structure (the second term) may lead to compensatory estimation of latent classes that differ in their class means (the first term).

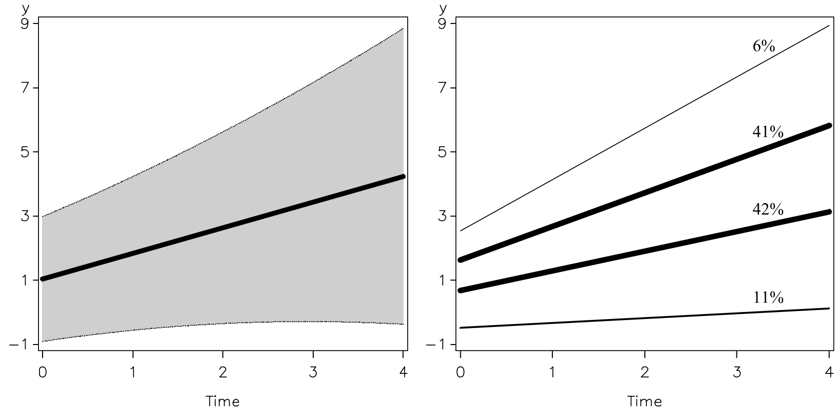

FIGURE 3 Left: Depiction of a latent curve model where the average trend is the solid line and variability is represented by the shaded region enclosing 95% of the individual trajectories at any given timepoint. Right: The class trajectories obtained from a four-class growth mixture model assuming no within-class variability.

As a salient example, consider the case where the data arises from a LCM (there is only one class, so all individual variability is within-class variability), but a Nagin-type GMM is fit to the data that constrains the within-class variability to be zero (e.g., $\mathbf{\Psi}_k = \mathbf{0}$) and the residuals to be uncorrelated (e.g., $\mathbf{\Theta}_k$ is diagonal). In this case, the misspecification of the within-class model forces the estimation of spurious latent classes that differ in their mean trajectories to capture the covariances among the repeated measures. Visually, the comparison between the true model and fitted model is shown in Figure 3. In the left panel, the true continuous distribution of individual trajectories is indicated by the 95% confidence bands, whereas in the right panel the individual variability is parsed into discrete group trajectories. These group trajectories do not represent "distinct" subgroups, they are simply a necessary consequence of the fact that the repeated measures are correlated. Similarly, other misspecifications of the covariance structure (other than assuming $\mathbf{\Psi}_k = \mathbf{0}$) could also lead to compensatory estimation of spurious latent classes.

As model specification checks are not yet well developed for GMMs, it is unclear how often, in practice, covariance structure misspecifications have impacted latent class estimation. For models involving a limited number of observed variables, selecting the number of classes based on fitting unrestricted normal mixture models may help to guard against this possibility (Bauer & Curran, 2004).

Assumption 3: Effects of exogenous predictors are linear. A key assumption of the conditional GMM is that the relationships between exogenous predictors and the individual trajectory parameters are linear within classes. This assumption was not explicitly considered in any of the 21 applications I reviewed. What, then, would be the consequence of failing to model a nonlinear effect? Bauer and Curran (2004) considered this question for mixtures of structural equation models and showed that the failure to model nonlinear effects can prompt the estimation of spurious latent classes.

To see how this problem might manifest in a GMM, I generated a sample of 600 cases from a conditional LCM that included a nonlinear effect of the exogenous predictor, x, on the intercepts of the individual trajectories, η_1. This nonlinear relationship is depicted in the top panel of Figure 4, superimposed on the actual intercept values for 200 of the simulated cases. Here, the intercept might represent initial levels of some sort of problem behavior, such as antisocial behavior, and x might be a protective factor, such as the quality of the home environment. Particularly low quality home environments might be predictive of higher initial levels of antisocial behavior, but there may be little difference in the antisocial behavior of children experiencing medium and high quality home environments, producing a curve of "diminishing returns."

I next fit a one-class GMM to this data, assuming that x is a linear predictor of η_1 (equivalent to a standard conditional LCM). This produced the regression line shown in the middle panel of Figure 4. The line is not completely misleading—it provides a first-order approximation to the true function and correctly indicates the negative relationship between the two variables—but it is also clearly an imperfect summary of the true relationship. If I now add a second class to the GMM, include x as a class predictor, and allow the within-class effect of x to differ across classes, I obtain the two regression lines shown in the bottom panel of Figure 4. This two-class model is preferred by Bayes' Information Criterion ($\mathrm{BIC}_2 = 9295 < \mathrm{BIC}_1 = 9414$).[5] Further, both the adjusted likelihood ratio test of Lo, Mendell, and Rubin (2001) and the boostrapped likelihood ratio test of McLachlan (1987) rejected the single-class model with $p < .0001$. As can be seen, the improvement in fit is due to the fact that each regression line captures a different aspect of the underlying nonlinear relationship. The regression line for the first class, characterized mainly by low x values, is steeply negative, whereas the regression line for the second class, characterized by medium and high x values, is virtually flat, reflecting the asymptotic nature of the relationship. We might thus profitably regard the two classes as providing local approximations to the true function (Bauer, 2005). In practice, however, it is more likely that the classes would be interpreted as two qualitatively distinct population subgroups, despite the fact that the individual differences are in fact quantitative in nature.

Whether unmodeled nonlinear effects produced spurious classes in the 21 applications I reviewed is unclear, but it is certainly a possibility. For the majority

[5] Across 500 simulated data sets, the two-class GMM was favored over the one-class GMM by the BIC 100% of the time.

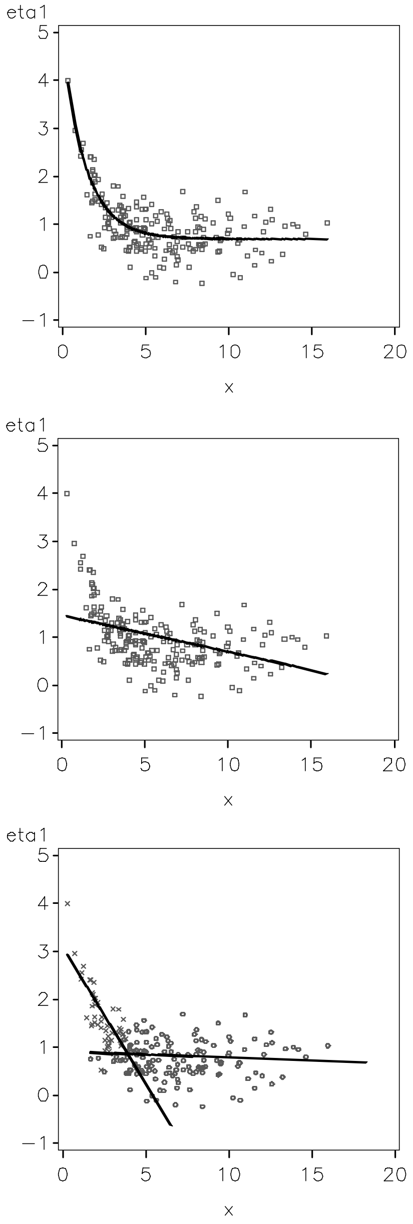

FIGURE 4 Top: The nonlinear relationship between predictor x and trajectory intercepts in the population. Middle: Result of fitting a latent curve model for which the effect of x was specified as linear. Bottom: Result of fitting a two-class growth mixture model for which the within-class effect of x was specified as linear (symbols indicate class membership based on modal class probability). Two hundred data points shown.

of the applications, the range of the dependent variable was limited and floor effects were likely. A nonlinear, asymptotic regression function, such as that depicted in the left panel of Figure 4, is often better justified statistically in such cases than a linear function that does not respect the limited range of the outcomes. In practice, this assumption can be checked by adapting standard regression diagnostics (e.g., plotting residual factor score estimates against predicted values from a conditional LCM).

Assumption 4: Missing data are MAR. As typically used, GMMs assume that the data are missing at random. In the applications I reviewed, the attrition rate averaged 20% (ranging from 3% to 44%). Given that most of these applications involved sensitive and sometimes criminal behaviors, missingness may be partially a consequence of the missing values, even after conditioning on the observed data, a violation of the MAR assumption. Further, when cases with completely missing data differ from those that have at least partial data, this results in the well-known problem of non-response bias. Given the mean consent rate of 77% (ranging from 54% to 90%), this too is a potentially serious issue for the applications I reviewed.

One reason for concern is that non-ignorable missingness can result in distorted distributions for the observed data. For instance, if the probability of missingness for y is an increasing function of y, then high y values will be under-represented in the observed data distribution, producing negative bias in the sample mean and variance estimates. Additionally, the overall shape of the observed data distribution may be altered relative to what would have been observed with complete data. If the distributions of the repeated measures are sufficiently distorted, this may in turn impact recovery of the correct number of latent classes in a GMM analysis (see Assumption 1).

As an initial evaluation of this issue, I conducted a small simulation study using three different population structures for five repeated measures: an LCM, a two-class GMM with equal class sizes, and a two-class GMM with unequal class sizes (details shown in Table 1). I then deleted a subset of the simulated data values to produce three missingness patterns: complete data, 23% non-response and 20% attrition (average missingness), or 42% non-response and 39% attrition (extreme missingness). The probability of missingness was an increasing function of the dependent variable, resulting in a clear violation of the MAR assumption. Five hundred replications with $N = 500$ cases contributing at least one non-missing observation were generated for each combination of population model and missingness pattern. GMMs with one, two, and three classes were then fit to the data and the best fitting model was determined by examining the BIC. At this sample size, the BIC has been shown to outperform the adjusted likelihood ratio test for selecting the number of classes in a growth mixture model and to perform as well as the more computationally intensive

TABLE 1
Population Parameter Values for Simulations Evaluating Influence of Non-Randomly Missing Data on the Estimation of Growth Mixture Models (GMMs).
Population Models Include a Linear LCM (1-Class GMM), and 2-Class Linear GMMs with Equal or Unequal Class Proportions

		GMM (50/50)		*GMM (25/75)*	
Parameter	*LCM*	*Class 1*	*Class 2*	*Class 1*	*Class 2*
α_1	5	6	4	6.5	4.5
α_2	1	1.2	.8	1.3	.9
ψ_{11}	2	1		1.25	
ψ_{21}	−.1	−.3		−.25	
ψ_{22}	.25	.21		.22	
$\theta_{11}, \theta_{22}, \theta_{33}, \theta_{44}, \theta_{55}$	.67, .68, .87, 1.22, 1.73				
Missing Data Process					
Average	Non-Response:	logit(missing y_1 to y_5) $= -6 + .87y_1$			
	Attrition ($t > 1$):	logit(missing y_t to y_5) $= -6 + .39y_t$			
Extreme	Non-Response:	logit(missing y_1 to y_5) $= -6 + 1.1y_1$			
	Attrition ($t > 1$):	logit(missing y_t to y_5) $= -6 + .51y_t$			

Note. $\Lambda = [1\ 0,\ 1\ 1,\ 1\ 2,\ 1\ 3,\ 1\ 4]$ for all models. Models are mean and covariance equivalent. Classes are equidistant in two GMMs.

bootstrapped likelihood ratio test (Nylund, Asparouhov & Muthén, 2007). As such, only the BIC was used for model selection in this demonstration.

When data were generated from the LCM, the single-class model was correctly selected 99.6% of the time with complete data, 99.8% of the time with average missingess, and 97.8% of the time with extreme missingness. Examination of the repeated measures distributions suggested that the missingness mechanism was not strong enough, even in the extreme condition, to induce sufficient non-normality that extra latent classes would be needed to reproduce the data. The two-class GMM with equal class sizes displayed similar robustness: the two-class model was correctly selected in over 99% of replications in all three missingness conditions. In this case, the mixture of the two classes produced repeated measures distributions with strong negative kurtosis (e.g., flat and sometimes bimodal peaks). This shape persisted even with the selective deletion of high values from the repeated measures distributions and hence model selection was unaffected by the missing data.

In contrast to the other two population models, the estimated latent class structure for the two-class GMM with unequal class sizes was less robust: the percentage of replications for which the two-class model was correctly selected decreased from 95.2% with complete data to 75.8% with average missingness to 40.2% with extreme missingness. When the two-class model was not preferred, the one-class model (LCM) was almost always selected.[6] This result can again be explained by considering how the observed data distributions of the repeated measures deviated from the complete data distributions. Given the mixture of a large "low" class and a small "high" class, the dominant characteristic of the complete data distributions was positive skew. Selectively deleting high values from these distributions reduced their skew, producing distributions that were more normal in appearance. Differentiating between one- and two-class models then became quite difficult.[7]

In summary, under the limited set of conditions considered here, it appears unlikely that violation of the MAR assumption would result in the over-extraction of latent classes if data were actually generated from a single-group LCM. In contrast, there is some risk that, when classes do exist, smaller extreme classes will be under-represented in the observed data and hence become more difficult to recover. Note, however, that even if the correct number of latent classes has been identified, the parameter estimates obtained from the fitted models can still be badly biased if data is non-ignorably missing. Two possible corrections for this problem are to use pattern mixture models (discussed later) or to adjust for non-response and attrition through the use of probability weights (see following).

Assumption 5: Sampled individuals are independent and self-weighting. When fitting a GMM, one usually assumes both that the individuals are independent and that the observations are self-weighting. Indeed, this assumption was made in each of the 21 applications I reviewed, primarily because it simplifies model estimation. If the data are independent, then the individual log-likelihoods can be summed to arrive at the overall log-likelihood for the model. Additionally, if the sample is self-weighting, then one needn't worry about incorporating probability weights in the analysis to arrive at accurate effect estimates for the population. Samples that involve unequal probabilities of selection are not self-weighting. This commonly occurs in complex sample designs that also feature stratification and/or clustering. For these samples, the sampling probabilities, stratification variables, and clustering typically must be incorporated

[6] To confirm that this trend was not due simply to the loss of information associated with missing data, the same patterns of missingness were produced via a completely random process. With the MAR assumption in tact, the two-class model was preferred in 94% and 91.2% of the replications given average and extreme amounts of missing data, respectively. Thus, it is nature and not just the extent of missing data that affects the estimated latent class structure.

[7] Some studies have found that a sample-size adjustment to the BIC proposed by Sclove (1987) improves selection of the number of classes for a GMM (Tofighi & Enders, 2007; see also Lubke & Neale, 2006), whereas others have shown superior performance for the unadjusted BIC (Nylund, Aparouhov, & Muthén, 2007). In the present simulation, the sample-size adjusted BIC was more likely to select too many classes than the BIC, and hence only the more conservative results for the BIC are reported.

into the analysis to generate accurate estimates for the target population. There are two main approaches for doing so—a design-based approach and a model-based approach (see B. O. Muthén & Satorra, 1995, for a contrast of the two approaches for structural equation models). Both of these approaches can be implemented with GMMs (L. K. Muthén & Muthén, 2007). For example, with the model-based approach, one could include random effects to account for clustering effects and/or incorporate stratification variables and selection variables as covariates.

None of the GMM applications I reviewed used these strategies. In fact, only a few of the samples could be described as probability samples at all. Most were community-based or clinical samples, for which the target population was loosely defined and the probabilities of selection unknown. Those samples that could be described as probability samples, were complex probability samples, but in no case were probability weights used in the analysis (nor was the complex design tested for informativeness to justify omitting the weights; Asparouhov, 2006). Additionally, several applications involved clustered data structures (e.g., saturation sampling of students at a particular grade level within multiple schools), but clustering effects were neither assessed nor modeled.

Little is known about the consequences of ignoring unequal probabilities of selection and/or clustering effects on latent class estimation in GMMs or, for that matter, finite mixture models in general. In one of the few papers considering this issue, Wedel, ter Hofstede, & Steenkamp (1998) showed that if the usual ML estimator (which assumes a self-weighting, independent sample) is applied to complex sample data, it can result in the selection of too many classes. Further, even if the correct number of classes is selected, the class proportions and within-class model estimates can be badly biased. As an alternative, they formulated a probability-weighted pseudolikelihood estimator and showed the results of this estimator to be more accurate. This estimator may also provide one way to correct for non-ignorably missing data processes, which may be viewed as producing unequal probabilities for data to be observed and included in the analysis.

We can thus conclude that, in practice, investigators are routinely fitting GMMs that assume independent, self-weighted data to complex probability samples and convenience samples, with the possible consequences that they are estimating the wrong number of classes and obtaining incorrect estimates within classes. When selection probabilities are known, using the pseudolikelihood estimator and incorporating stratification variables and clustering effects into the models should ameliorate these concerns.

Summary of Methodological Concerns

Overall, the portrait that emerges from this review of contemporary GMM applications suggests that the assumptions of the model are infrequently checked and rarely satisfied. This should come as no surprise, as the assumptions of very few models are met in practice. For instance, the assumptions of the LCM are probably violated equally often. However, it seems that the GMM is less robust than the LCM (and many other models). Specifically, an incorrect latent class structure may result from simple non-normality, misspecification of the covariance structure, failure to model nonlinear effects, violation of the MAR assumption for missing data, or insufficient attention to the sampling design. With the notable exception of the MAR assumption, the result is typically the estimation of too many classes. The tendency for too many classes to be estimated when assumptions of the GMM are violated is disturbing. The classes are imbued with such importance—thought to reflect distinct etiologies, differential risk, and targets for intervention—that getting the classes "wrong" is a serious matter.

The methodological challenges identified here are not necessarily intractable, but addressing them would require massive changes in the way psychology research is done. Data would need to be collected on probability samples with sufficiently precise measurement to justify the conditional normality assumption. Even then, there would need to be a strong rationale for why the observations within classes *ought* to be normally distributed (perhaps based on prior research with more homogeneous samples). Alternatively, different assumptions could be made about the distributions of the repeated measures (e.g., with ordinal or count data), though this would require equally strong justification. Diagnostics would need to be conducted to evaluate possible nonlinear relationships, and missing data would need to be minimized (or, at least, the MAR assumption should be reasonable). Finally, some way of guarding against or detecting misspecifications of the within-class mean and covariance structure would be required (e.g., determining the number of classes first by fitting unrestricted mixture models, then proceeding to the GMM). The recent development of goodness of fit indexes, such as the skew and kurtosis tests of B. Muthén (2003), or the component property method of Hellemann (2006), may also help to differentiate between models in which the classes are spurious versus those in which the classes are at least plausibly "real" (though see Bauer & Curran, 2003b, for a somewhat more skeptical view).

Even if these methodological challenges can be met, however, there are more fundamental reasons to question the use of GMMs in psychological research, to which we now turn.

THEORETICAL CONCERNS

Earlier, I noted that when fitting a model, the core assumptions of the model should be met (within the degree of error permitted by robustness conditions), the model should be a reasonable approximation to reality, and the results should be scientifically meaningful. We have seen that, very often, the first of these criteria does not hold for GMM applications—the assumptions of the GMM are typically grossly violated, with the likely consequence of producing a latent class structure that bears little resemblance to reality. However, as indicated previously, these assumptions could, in principle, be met or changed. Might the GMM then offer a reasonable approximation to reality?

The best GMM applications are motivated by strong taxonomic theories and include construct validation analyses to increase confidence in the obtained trajectory classes (e.g., Odgers et al., 2007). Often, however, there is little theoretical justification for the existence of discrete groups. The implicit motivation seems to be that we cannot expect our populations to be homogeneous and hence should favor GMMs over LCMs. LCMs do not, however, assume that change over time is homogeneous in the population. Rather, each individual is accorded his or her own personal trajectory. Thus, both models allow for heterogeneity in change over time, they just make different assumptions about how this heterogeneity is distributed.

Another theoretical vaguery that is often employed to motivate the application of GMMs is that the models are "person-centered" as opposed to "variable-centered" (Connell & Frye, 2006; Muthén & Muthén, 2000; Nagin, 2005, p. 15). This claim conflates methodology with theory. The person-centered (or person-oriented) approach to psychological research is deeply rooted in the holistic-interactionist paradigm articulated by Magnusson & Törestad (1993). The key idea is that psychological research should study the person as a whole rather than the cumulative effects of individual variables. Methodologically, this often translates into using cluster analysis to identify patterns across a set of variables. The focus is then on these holistic patterns, not the individual variables that went into the analysis.

The claim that GMMs are person centered seems to stem from the fact that individuals are clustered. But the clustering is done for an entirely different reason. In a GMM, latent classes do not represent coherent patterns within an interactive multivariable system. They are instead subgroups defined by change in a single variable measured repeatedly through time. Given this, Bergman and Trost (2006) noted that GMMs can be considered person centered only if this single variable represents all of the key aspects of the system under study. From the perspective of the holistic-interactionist paradigm, it is rather hard to imagine this being possible. Bauer & Shanahan (2007) hence concluded that referring to GMMs as person centered is misleading. If what is meant by the incorrect use of this term is that the group trajectories define types of people, then it is worth repeating that each person has his or her own unique trajectory in the LCM as well, making it equally "person centered."

Dispensing with these two pseudotheoretical rationales for fitting GMMs, we are faced with the more fundamental question of whether theory predicts the presence of a latent taxonomy consisting of qualitatively distinct groups, as in Moffitt (1993), or whether we should instead expect individuals to differ by gradations. Given that most behaviors, personality traits, and abilities are multi-determined, involving numerous epigenetic inputs, it seems reasonable to expect that individual trajectories will most often differ continuously, by degrees, rather than by types.[8] As Maughan (2005) states, "There is ... widespread recognition that many of the behaviors we study are dimensionally distributed and do not show clear-cut points differentiating 'normality' and 'pathology'" (pp. 120–121). Nagin (2005) agrees, noting that "although there may be populations made up of groups that are literally distinct, they are not the norm. Most populations are composed of a collection of individual-level developmental trajectories that are continuously distributed across population members" (p. 45).

By this reasoning, latent trajectory classes are, in most cases, nothing more than artificial categories rendered from the true continuum of change. And these categories can be fickle. Change the start values for the estimator, add covariates, relax or impose a few constraints, collect one more wave of data, or alter your measurement instrument and wholly new categories may emerge, none any more or less valid than the original (Eggleston et al., 2004; Hipp & Bauer, 2006; Jackson & Sher, 2005, 2006). In general, we should be wary of models that are highly sensitive to small changes in the data or model specification. Further, categorization of continuous variability can make inference difficult, if not outright hazardous (Bauer & Curran, 2003a). Are the latent classes then still scientifically useful, as some have argued?

Groups That Do Not Exist

Those who advocate using GMMs even in the absence of taxonic hypotheses make three points in support of their position. First, GMMs can capture a wide variety of distributions and hence provide a way to model non-normal random effects within a growth model (Nagin, 2005; B. Muthén & Asparouhov, in press; Segawa, Ngwe, Li, Flay, & Aban Aya Coinvestigators, 2005; Vermunt & van Dijk, 2001). This is undeniably true; however, it is also true that the estimates obtained from conventional growth models are robust to the presence of non-normal random effects (Raudenbush & Bryk, 2002, p. 274; Verbeke & LeSaffre, 1997). To be fair, this robustness does not necessarily extend to models for discrete outcomes, for which it may be more important to account for non-normality of the random effects. Even then, however, mixture distributions are not the only answer. Other approaches for semiparametrically modeling the

[8] A case can sometimes be made for types even when behaviors are multidetermined, for instance by arguing that only certain developmental pathways represent coherent patterns of adaptation (or maladaptation, as the case may be) or that threshold effects trigger developmental bifurcations.

continuous (but not necessarily normal) distribution of the individual trajectories exist that do not resolve the population into artificial groups (e.g., Chen, Zhang, & Davidian, 2002; Zhang & Davidian, 2001).

The danger of artificial groups is the potential for reification by scientists, policy makers, and practitioners. Statements like, "X% of the population belongs to group Y" and, "The odds of being in group Y relative to group Z increase twofold if you have characteristic W," have little apparent meaning if the groups are mere figments of the analysis and unstable from one study to the next. As Raudenbush (2005) observed, "Perhaps we are better off assuming continuously varied growth a priori and therefore never tempting our audience to believe in the key misconception that groups of persons actually exist. We would then not have to warn them strongly against 'reification' of the model they have been painstakingly convinced to adopt" (p. 136).

A second reason sometimes offered for using GMMs in the absence of true population subgroups is to highlight unusual trajectories that could easily be overlooked in an aggregated analysis. For instance, by focusing on expectations in a conventional growth model, one might fail to appreciate more extreme or infrequent trajectories that could be revealed by a GMM (Collins, 2006; Nagin & Tremblay, 2005). It is no doubt true that GMMs can be used to help visualize the variability of individual trajectories, including variability that might not be predictable on the basis of observed covariates. The picture that results is easy to interpret, substantively compelling, even "seductive" (Sampson & Laub, 2005). But a plot of four or five trajectories necessarily represents a gross simplification of the full variability in the population. Moreover, we need not use a mixture model to identify unusual trajectories. Any growth analysis should begin with a graphical assessment of the individual trajectories to identify both typical and unusual patterns of change, no artificial groups required (Hedeker & Gibbons, 2006, p. 54; Singer & Willett, 2003, pp. 24–35).

Finally, a third argument sometimes offered in favor of GMMs is that accurate inferences can still be drawn from the models even when the groups are admittedly fictional (Nagin, 2005). The estimates obtained for the groups are interpreted to reflect regional conditions in the globally continuous distribution of individual change. Bauer & Shanahan (2007) made much the same interpretation of the mixing components within a latent profile analysis, but they also remarked that relatively little research has yet been conducted to see how well the properties of the global distribution are recovered via this local approximation (see Brame, Nagin, & Wasserman, 2006; and B. Muthén & Asparouhov, in press, for initial studies in the GMM context). More fundamentally, by resolving the population into latent subsets, one runs two risks. First, power is diminished. This was shown by simulation in Bauer & Curran (2003a) and has also been seen in real-data comparisons of GMMs relative to standard growth models (NICHD Early Child Care Research Network, 2004, p. 100). Second, within the general GMM, Bauer & Curran (2003a) detected spurious class × covariate interactions at a relatively high rate. B. Muthén (2004) provides an explanation of why this may occur: "[A] GGMM that incorrectly divides up trajectories in, say, low, medium and high classes might find that the covariates have lower and insignificant influence in the low class due to selection on the dependent variable" (p. 253). Note that these risks also attend other methods of categorizing continua (MacCallum, Zhang, Preacher, & Rucker, 2002).

To summarize, although I do not disagree with any of these arguments concerning the potential validity of GMMs even in the absence of a taxonomic theory, I see the cost-benefit ratio much differently. The potential risks, which include reification, over-simplification, lost power, and spurious effects, seem to me to be much greater than the possible benefits. To be clear, I do think there are some situations in which groups exist, or at least putatively exist, for which mixture analyses may be appropriate and valuable. More often, however, the groups appear to serve only as a simplification of continuous variability, leading to the fundamental question of whether group-based inferences are sensible in the absence of real groups. In the next section, I describe a possible alternative way to use GMMs that does not involve this apparent contradiction.

Whole-Population Inference With Growth Mixture Models

In an influential monograph on finite mixture modeling, Titterington, Smith, and Makov (1985, pp. 2–3) distinguished between two quite different types of applications. In *direct applications*, the latent classes are interpreted as representative of distinct population subgroups. In contrast, in *indirect applications*, the latent classes are used solely to provide a tractable form of analysis for data that may not obey traditional parametric models. To date, most interest in GMMs has been centered on direct applications and specifically the evaluation (or generation) of taxonomic theories of psychosocial development. In contrast, those who argue for the use of GMMs even when no groups are thought to exist are promoting indirect applications of the model. We may further subdivide indirect applications into two types. In the first, the latent class estimates are presented and interpreted similarly to a direct application, but with the caveat that the groups are heuristic, not real (Nagin, 2005). The second type of indirect application, with which I am more comfortable, instead places the focus for interpretation on the overall distribution obtained from the mixture, not the component distributions of the latent classes.

To clarify these distinctions, let us reconsider the results presented in Figure 4. Recall that the top panel depicts the true, continuous nonlinear relationship between the predictor x and the intercepts of the individual growth trajectories in the population. In contrast, the bottom panel depicts the linear relationship estimated for each of two latent classes. If the model has been fit as a direct application, then the latent classes are spurious, and interpreting them as fundamentally different population subgroups would be incorrect and potentially misleading. One might, however, still justify fitting the model as an indirect application of the first type. Following the recommendation of Nagin (2005), we could present the groups as a heuristic device and validly interpret the within-class effect estimates as reflecting local conditions. That is, within the region of the distribution characterized by especially low values of x, the effect of x is steeply negative (Class 1); within the remaining region, x has little or no effect on the trajectory intercepts (Class 2). This interpretation is perfectly reasonable, but it continues to place emphasis on the groups rather than the population as a whole. Because the groups are admittedly artificial, it seems odd to make them the focus of interpretation. Moreover, the problems noted earlier continue to arise—there is the potential for reification (e.g., x represents a distinct etiological mechanism associated with Class 1 but not Class 2), and the power to detect the overall effect of x on the trajectory intercepts may be diminished by parsing the total sample into smaller subgroups.

Given the artificial nature of the groups, it would seem logical to return to the level of the whole population when making inferences, as in the second type of indirect application. To do so, we must compute the expected value of the trajectory intercepts mixed over classes. Bauer (2005) demonstrated one way that this can be done when the predictor is latent. In this case, however, the predictor is observed so a simpler approach will suffice. Specifically, the expected value we require may be computed from the general equation

$$E(\boldsymbol{\eta}_i|\mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i)\,(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i)\,, \tag{12}$$

where $\pi_{ik}(\mathbf{x}_i)$ is defined as in Equation 10. As can be seen, the mixed effect estimate is a weighted sum of the within-class predicted values, where the weights are the probabilities of class membership given $\mathbf{x}$. Applying this result to the example data, we obtain the curve depicted in Figure 5, which can be seen to reproduce the true population function shown in the top panel of Figure 4 quite well.

It is worth noting that Figure 5 was produced by fitting the very same model and using the same parameter estimates that produced the bottom panel of Figure 4. The depiction in Figure 4, however, requires us to make conditional inferences—the effect of x is strong in one group and weak in the other—despite the expressly fictional status of the groups. In contrast, in Figure 5, we capitalize on the local approximations afforded by the two latent classes to reconstitute the global relationship between the two variables. Our inferences are no longer conditional on group membership but rather pertain to the population as a whole. We have thus avoided the apparent logical inconsistency of

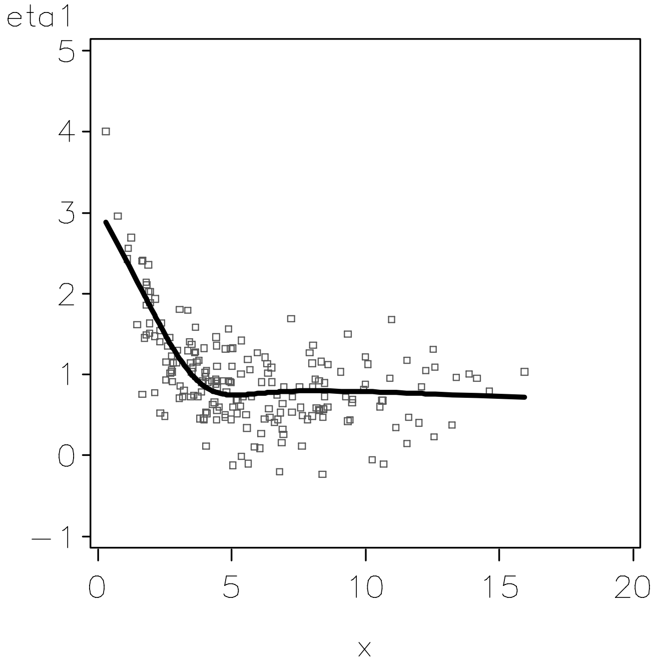

FIGURE 5 Approximation of the nonlinear effect in the top panel of Figure 4 via an indirect application of a growth mixture model (GMM). Two hundred data points shown.

making inferences about groups while at the same time denying their existence. Other key advantages of focusing on the total population include the preservation of continuity, diminished temptation for reification, and potentially greater power.

Given my own research on the topic, I have naturally focused here on nonlinear effect estimation, but there are many other ways that indirect applications of GMMs could be useful. For instance, another interesting indirect application is the latent pattern mixture model (LPMM) of Roy (2003) and Lin, McCulloch, and Rosenheck (2004). The LPMM represents an extension of the pattern mixture approach to the analysis of non-ignorably missing data. In a classical pattern mixture analysis, the sample is first stratified by observed missingness patterns, then key effects (e.g., time trends, or time by covariate interactions) are estimated for each pattern and, finally, the estimates are averaged over strata to obtain the mixed, or aggregate, effect estimates (Hedeker & Gibbons, 1997; Little, 1995; see also related work by Allison, 1987, and McArdle & Hamagami, 1992). The advantage of the pattern mixture model is that it provides accurate estimates even when data are non-ignorably missing. Unfortunately, pattern mixture models become more difficult to implement as the number of missingness patterns increases (e.g., with long time series and intermittent missingness). The LPMM offers a solution to this problem by replacing the large number of observed missingness patterns with a smaller set of latent missingness patterns (Lin et al., 2004; Morgan-Lopez & Fals-Stewart, 2007; Roy, 2003). The latent missingness patterns are jointly defined by binary missing data indicators for each timepoint as well as the observed repeated measures. Similar to a standard pattern mixture analysis, the within-class estimates are averaged across the latent missingness patterns to produce estimates for the aggregate-level effects.

What both of these example holds in common is the use of latent classes as a convenient statistical device to better enable inference at the level of the whole population, avoiding the problematic interpretation of non-existent groups.

CONCLUSION

GMMs are being used in psychology at an increasing rate. The fundamental question I sought to address here is whether these models are likely to advance psychological science. My firm conviction is that, if these models continue to be applied as they have been so far, the answer is clearly no. Most current applications of GMMs cannot help but find multiple trajectory classes, given the ubiquity of assumption violations and lack of robustness of the models. This results in a strong confirmation bias when GMMs are used to evaluate non-specific taxonic conjectures. In contrast, the evaluation of specific taxonic hypotheses will often be frustrated by the addition of spurious groups and/or biased estimates for the trajectory classes. I therefore believe that direct applications of GMMs should be refrained from unless both the theory and data behind the analysis are uncommonly strong. Otherwise, the application of GMMs in psychological research is likely to lead to more blind alleys than ways forward.

An alternative place for GMMs in psychological research is as an approximating device in indirect applications. Concerns about spurious classes then no longer apply, given that the classes are expressly acknowledged to be artificial. When used in this way, a key consideration is whether to continue to interpret the latent classes, despite the temptation for reification. My own view is that it is preferable to reintegrate the results prior to interpretation. The idea is to capitalize on the flexibility of the latent classes to capture features of the data that ordinary growth models may miss while at the same time avoiding the problematic interpretation of fictional groups. An interesting question, however, is whether GMMs will actually be used for this purpose when the idea of “real” groups seems to be so much more alluring.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation grant SES-0716555. I thank Patrick Curran, Robert MacCallum, and Sonya Sterba for helpful comments on an earlier draft of this article. I also thank Kenneth Bollen and Patrick Curran for their invaluable advice and guidance during and after my post-doctoral training. Much of the research upon which this article was based would not have been possible without their support and active collaboration.

REFERENCES

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology 1987*, 71–103.

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

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods, 35*, 439–460.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling: A Multidisciplinary Journal, 4*, 513–535.

Bauer, D. J., & Curran, P. J. (2003a). 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. (2003b). Over-extraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods, 8*, 384–393.

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.

Bauer, D. J., & Shanahan, M. J. (2007). Modeling complex interactions: Person-centered and variable-centered approaches. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 255–283). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bergman, L. R., & Trost, K. (2006). The person-oriented versus the variable-oriented approach: Are they complementary, opposites, or exploring different worlds? *Merrill-Palmer Quarterly, 52*, 601–632.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach.* Hoboken, NJ: Wiley.

Brame, R., Nagin, D. S., & Wasserman, L. (2006). Exploring some analytical characteristics of finite mixture models. *Journal of Quantitative Criminology, 22*, 31–59.

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

Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions 2* (pp. 171–198). Amsterdam: North-Holland.

Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology, 41*, 193–208.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147–158.

Chen, J., Zhang, D., & Davidian, M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics, 3*, 347–360.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Connell, A. M., & Frye, A. A. (2006). Growth mixture modelling in developmental psychology: Overview and demonstration of heterogeneity in developmental trajectories of adolescent antisocial behavior. *Infant and Child Development, 15*, 609–621.

Eggleston, E. P., Laub, J. H., & Sampson, R. J. (2004). Methodological sensitivities to latent class analysis of long-term criminal trajectories. *Journal of Quantitative Criminology, 20*, 1–26.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*, 64–78.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Hellemann, G. S. (2006). *The component property method: A new approach to testing the number of components in a finite mixture model*. Unpublished doctoral dissertation, UCLA.

Hipp, J., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods, 11*, 36–53.

Hoeksma, J. B., & Kelderman, H. (2006). On growth curves and mixture models. *Infant and Child Development, 15*, 627–634.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors, 19*, 339–351.

Jackson, K. M., & Sher, K. J. (2006). Comparison of longitudinal phenotypes based on number and timing of assessments: A systematic comparison of trajectory approaches II. *Psychology of Addictive Behaviors, 20*, 373–384.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York: Academic Press.

Land, D. K., & Nagin, D. S. (1996). Micro-models of criminal careers: A synthesis of the criminal careers and life course approaches via semiparametric mixed Poisson models with empirical applications. *Journal of Quantitative Criminology, 12*, 163–191.

Lin, H., McCulloch, C. E., & Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics, 60*, 295–305.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112–1121.

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

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41*, 499–532.

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.

Magnusson, D., & Törestad, B. (1993). A holistic view of personality: A model revisited. *Annual Review of Psychology, 44*, 427–452.

Maughan, B. (2005). Developmental trajectory modeling: A view from developmental psychopathology. *The Annals of the American Academy of Political and Social Science, 602*, 119–130.

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

McArdle, J. J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research, 18*, 145–166.

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

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.

Morgan-Lopez, A. A., & Fals-Stewart, W. (2007). Analytic methods for modeling longitudinal data from rolling therapy groups with membership turnover. *Journal of Consulting and Clinical Psychology, 75*, 580–593.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods, 8*, 369–377.

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

Muthén, B., & Asparouhov, T. (in press). Growth mixture analysis: Models with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Advances in longitudinal data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

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

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., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological Methodology 1995*, 216–316.

Muthén, L. K., and Muthén, B. O. (2007). *Mplus User's Guide.* (4th ed.). Los Angeles: Author.

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

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

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

Nagin, D. S., & Tremblay, R. E. (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. S., & Tremblay, R. E. (2005). Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology, 43*, 873–904.

NICHD Early Child Care Research Network (2004). Trajectories of physical aggression from toddlerhood to middle childhood. *Monographs of the Society for Research in Child Development, 69*(4, Serial No. 278).

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

Odgers, C. L., Caspi, A., Broadbent, J. M., Dickson, N., Hancox, R. J., Harrington, H., et al. (2007). Prediction of differential adult health burden by conduct problem subtypes in males. *Archives of General Psychiatry, 64*, 476–484.

Raudenbush, S. W. (2005). How do we study "What happens next"? *The Annals of the American Academy of Political and Social Science, 602*, 131–144.

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

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent droupout class model. *Biometrics, 59*, 829–836.

Sampson, R. J., & Laub, J. H. (2005). Seductions of method: Rejoinder to Nagin and Tremblay's "Developmental trajectory groups: fact or fiction?" *Criminology, 43*, 905–913.

Sampson, R. J., Laub, J. H., & Eggleston, E. P. (2004). On the robustness and validity of groups. *Journal of Quantitative Criminology, 20*, 37–42.

Sattora, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity, 24*, 367–386.

Sattora, A., & Bentler, P. M. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics and Data Analysis, 10*, 235–249.

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.

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

Segawa, E., Ngwe, J. E., Li, Y., Flay, B. R., & Aban Aya Coinvestigators (2005). Evaluation of the effects of the Aban Aya Youth Project in reducing violence among African American males using latent class growth mixture modeling techniques. *Evaluation Review, 29*, 128–148.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

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

Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in a growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.

Verbeke, G., & LeSaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis, 23*, 541–556.

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.

Wedel, M., ter Hofstede, F., & Steenkamp, J.-B. E. M. (1998). Mixture model analysis of complex samples. *Journal of Classification, 15*, 225–244.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika, 69*, 421–436.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*, 795–802.