Multivariate Behavioral Research, 47:590–634, 2012

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.692639

Psychology Press
Taylor & Francis Group

Factors Affecting the Adequacy and Preferability of Semiparametric Groups-Based Approximations of Continuous Growth Trajectories

Sonya K. Sterba
Vanderbilt University

Ruth E. Baldasaro and Daniel J. Bauer
The University of North Carolina at Chapel Hill

Psychologists have long been interested in characterizing individual differences in change over time. It is often plausible to assume that the distribution of these individual differences is continuous in nature, yet theory is seldom so specific as to designate its parametric form (e.g., normal). Semiparametric groups-based trajectory models (SPGMs) were thus developed to provide a discrete approximation for continuously distributed growth of unknown form. Previous research has demonstrated the adequacy of the approximation provided by SPGM but only under relatively narrow, theoretically optimal conditions. Under alternative conditions, which may be more common in practice (e.g., higher dimension random effects, smaller sample sizes), this study shows that approximation adequacy can suffer. Furthermore, this study also evaluates whether SPGM's discrete approximation is preferable to a parametric trajectory model that assumes normally distributed random effects when in fact the distribution is modestly nonnormal. The answer is shown to depend on distributional characteristics of both repeated measures (binary or continuous) and random effects (bimodal or skewed). Implications for practice are discussed in light of empirical examples on externalizing behavior.

Correspondence concerning this article should be addressed to Sonya K. Sterba, Department of Psychology & Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203. E-mail: Sonya.Sterba@vanderbilt.edu

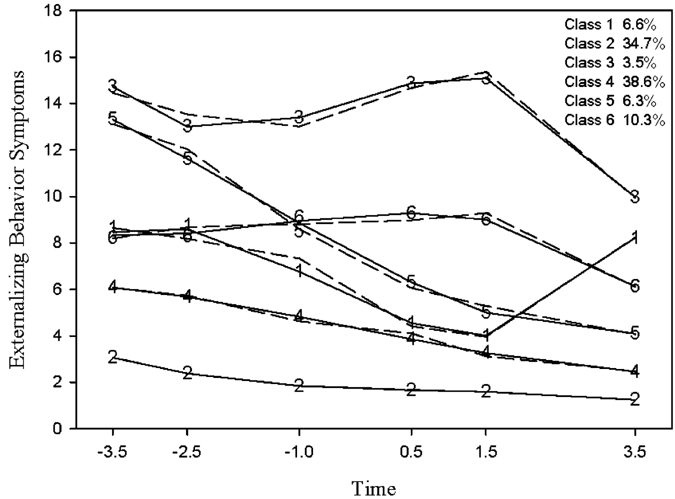

FIGURE 1 Semiparametric groups-based trajectory model of externalizing behavior problems: Six-class empirical example. *Note.* Time = (age in years—5.5). Solid lines = model estimated mean per class. Dashed lines = observed mean per class, calculated using modal class assignment. Proportions associated with each class are listed in the figure legend. Illustrative class labels: 1 = "decreaser/increaser"; 2 = "low-stable"; 3 = "chronic"; 4 = "moderate declining"; 5 = "high declining"; 6 = "moderate stable." Conditional response distribution assumed normal. Details on the empirical example are given in the section of this article titled Empirical Illustrations.

Characterizing individual differences in change over time is a core enterprise of psychology research. One data-analytic approach that can aid in this endeavor is the semiparametric groups-based trajectory model (SPGM; Nagin, 1999, 2005; Nagin & Land, 1993; also called a latent class growth model). This longitudinal method distills individual variation in change over time into a small set of discrete groups (or classes) of trajectories. Attractive to social science researchers, SPGM has been applied hundreds of times in psychology and related disciplines over the past decade.

To illustrate, consider an application of SPGM to repeated measures on externalizing behavior—the most common topic among 100 psychology SPGM applications we surveyed.[1] In Figure 1 (based on Sterba, Prinstein, Bauer, & Cox, 2005), individual differences in the course of externalizing are captured by six class-specific trajectories, which can differ in initial status, rate of change, and

[1]Out of 100 applications, 22 modeled externalizing or related antisocial, conduct, or aggressive behavior. See Online Appendix at http://www.vanderbilt.edu/peabody/sterba/appxs.htm for references for surveyed applications.

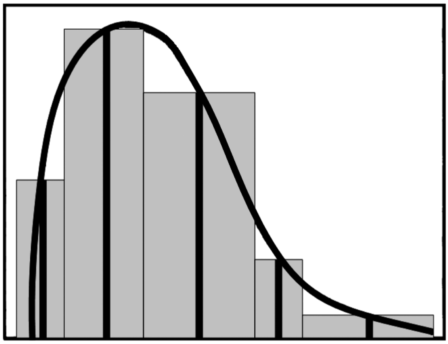

FIGURE 2 Heuristic sketch often cited in semiparametric groups-based trajectory model literature: Hypothesized semiparametric approximation of the probability density function of a single dimension random effect distribution. *Note.* Locations of points of support correspond to class means; heights of points of support correspond to class probabilities; number of points corresponds to number of classes.

class membership proportions. Researchers often desire to *directly* interpret such trajectory classes as literal or true latent subpopulations—where nature's joints have been carved (Titterington, Smith, & Makov, 1985). However, SPGM has been strongly motivated as a tool for providing a flexible, *indirect* semiparametric approximation of continuous individual differences in growth (e.g., Nagin, 2005; Nagin & Tremblay, 2005a,b,c,d). As summarized by Nagin and Tremblay (2001),

> As discussed in Nagin (1999), Nagin and Land (1993), and Nagin and Tremblay (1999), the assumption that the population is composed of distinct groups is unlikely to be strictly correct. Instead, the groups are intended as an approximation of an underlying continuous process. In so doing, we adopt a standard procedure in nonparametric and semiparametric statistics of approximating a continuous distribution from a discrete mixture. (p. 28)

To illustrate this role, methodological sources introducing SPGM routinely present a diagram resembling Figure 2 (e.g., Feldman, Masyn, & Conger, 2009; Muthén, 2004, 2008; Nagin, 1999, 2004, 2005; Nagin & Land, 1993; Nagin & Tremblay, 2005c; see also Sampson & Laub, 2005). The smooth curve depicts a one-dimensional distribution of individual differences in growth (say, differences in initial level, or intercept) in the population. An SPGM would extract $k = 1 \ldots K$ classes, which would serve as points of support for this continuous individual difference distribution, much like histobars in a histogram (here $K = 5$). The location of the point of support for the kth class is determined by a class-specific growth coefficient (here, class-specific intercept). The proportion of individuals in class k determines the height (or mass) of the kth point of support.

Together, these masses and locations define a discrete probability distribution for the SPGM.

Our literature review of 100 psychology SPGM applications (see Online Appendix at http://www.vanderbilt.edu/peabody/sterba/appxs.htm for random selection procedures) indicated that the ability of SPGM to indirectly semiparametrically approximate continuous individual differences in trajectories is widely referenced in empirical studies (e.g., Delucchi, Matzger, & Weisner, 2004; Kreuter & Muthén, 2008; Losoya et al., 2008; Louvet, Gaudreau, Menaut, Genty, & Deneuve, 2007; Maggi, Hertzman, & Vaillancourt, 2007; Mazza, Fleming, Abbott, Haggerty, & Catalona, 2009; Murphy, Brecht, Herbeck, & Huang, 2009; Nash & Kim, 2007; Obradovic, Burt, & Masten, 2006; Piquero, Fagan, Mulvey, Steinberg, & Odgers, 2005; Rodriguez-Zas, Southey, Whitfield, & Robinson, 2006; Segawa, Ngwe, Li, Flay, & Coinvestigators, 2005; van Ryzin, Chatham, Kryzer, Kertes, & Gunnar, 2009; Wiesner & Kim, 2006; Xie, McHugo, He, & Drake, 2010). However, whereas a direct interpretation of classes is intuitive, an indirect semiparametric function of classes is less so. Applications typically mention the indirect approximation function of classes only abstractly, without concrete connection to the obtained results or consideration of whether such a semiparametric approximation will be adequate or even necessary for their data. Hence, users of SPGM need to consider the following: first, whether this approximation will be adequate under conditions typical of psychological research, and second, whether it will be preferable to existing parametric models when the latter's distributional assumptions do not hold.

Regarding the first question, prior investigations of SPGM's semiparametric approximation function (Brame, Nagin, & Wasserman, 2006; Muthén & Asparouhov, 2008, pp. 158–161; Nagin, 2005, Chapter 3) considered very limited settings (e.g., low dimension distributions of individual differences and large sample sizes) that do not necessarily represent conditions widely seen in psychology studies. Further, there is theoretical reason to believe that the adequacy of SPGM's approximation will suffer in the opposite settings (higher dimension individual difference distributions and smaller N), as discussed later. Hence, our first goal is to consider more generally *when* SPGM's semiparametric approximation abilities will be adequate.

Regarding the second question, existing parametric growth models, such as hierarchical linear models (HLMs; also known as mixed effects models; Bryk & Raudenbush, 1987; Goldstein, 1986; Laird & Ware, 1982) also portray change in continuously distributed individual trajectories—but with the added requirement that growth parameters are multivariate normal. Using SPGM to discretely approximate such distributions imposes no such assumption. Furthermore, this assumption may not always be realistic; there is some indication that individual difference distributions for familiar psychological constructs may often be modestly nonnormal (e.g., van den Oord, Pickles, & Waldman, 2003). SPGM has been recommended when individual differences in growth potentially depart from normality (e.g., Nagin, 2004, 2005; Nagin & Tremblay, 1999, 2001, 2005a, 2005c). However, only one study has compared the two methods under this circumstance (Muthén & Asparouhov, 2008) and did so for highly nonnormal individual difference distributions. The second goal of this study is to extend the previous comparison by providing specifics on the relative performance of HLM and SPGM under modestly nonnormal continuous individual differences. As discussed later, theory suggests the categorical or continuous nature of the *repeated measures* will be an important distinguishing factor.

In sum, though the ability to semiparametrically approximate continuously distributed trajectories is a widely referenced motivation for SPGM, the quality of the approximation under different real-world circumstances and its performance relative to a potentially misspecified HLM has not been subject to sufficient scrutiny. This article aims to fill these gaps. First, we review the SPGM and HLM. Second, we review prior literature and provide theoretical justification for several hypotheses that are then tested via simulation. Implications of the simulation for interpreting SPGM results are discussed in the context of an externalizing behavior example.

ALTERNATIVE MODELS OF INDIVIDUAL CHANGE

Here we define the SPGM and HLM. Either model can be expressed for a variety of different distributions for the repeated measures (e.g., normal, binary [Bernoulli], count [Poisson]). (Conventionally the abbreviation HLM is used for normal outcomes and HGLM—hierarchical generalized linear model [McCulloch, Searle, & Neuhaus, 2008]—for discrete outcomes). Because SPGMs with count outcomes have been most commonly featured in SPGM simulations (e.g., Brame et al., 2006; Nagin, 2005) and discussions (e.g., Eggleston, Laub, & Sampson, 2004; Kreuter & Muthén, 2008; Nagin, 2004; Sampson & Laub, 2003; Sampson, Laub, & Eggleston, 2004) to date, we cover the normal and binary versions of SPGM and HLM/HGLM to better generalize to settings where, for instance, symptom levels or psychiatric diagnoses are being measured. To unify our model descriptions, we describe each in terms of a *conditional response distribution*, or probability distribution for the repeated measures conditional on predictors; a *linear predictor*, or linear combination of predictor(s); and a *link function* that transforms the range of the linear predictor to the range of the expected value of the conditional response distribution. In descriptions that follow, designate the outcome for person i at time t as y_{ti} where $t = 1 \ldots T$. Let $\mathbf{x}_{ti}$ denote a $p \times 1$ vector of predictor values at time t, which have fixed effects contained in the $p \times 1$ vector $\boldsymbol{\gamma}$. For instance $\mathbf{x}_{ti}$ might contain a 1 to define a trajectory intercept and a $time_{ti}$ score to capture change in y_{ti} (but could be expanded).

Semiparametric Groups-Based Trajectory Model

In the SPGM (Nagin & Land, 1993), the *conditional response distribution* of the repeated measures is conditional on individuals' scores on predictor(s) (e.g., time) and on individuals' latent class membership, where classes of a categorical latent variable c_i are indexed $k = 1 \ldots K$. For normal repeated measures $y_{ti}|\mathbf{x}_{ti}$, $c_i = k \sim N(\mu_{ti}, \sigma^2)$ and for binary repeated measures $y_{ti}|\mathbf{x}_{ti}$, $c_i = k \sim$ $BERNOULLI(\mu_{ti})$. Further, the *link function* for normal repeated measures is the identity $\eta_{ti} = \mu_{ti}$. The *link function* for binary repeated measures is typically logit $\eta_{ti} = \ln(\mu_{ti}/(1 - \mu_{ti}))$ or probit. In the SPGM, the *linear predictor* is a linear combination of predictor(s) for persons in the *k*th latent class:

$$\eta_{ti} = \mathbf{x}'_{ti}\boldsymbol{\gamma}^{(k)} \text{ if } c_i = k \tag{1}$$

where

$$P(c_i = k) = \frac{\exp(\delta_0^{(k)})}{\sum_{k=1}^{K} \exp(\delta_0^{(k)})}. \tag{2}$$

Equation (1) shows that in the SPGM, no systematic individual differences are allowed within class *k* because no random effects are included. This implies that persons within each homogeneous class *k* differ only due to random perturbations. Although fixed effects in $\boldsymbol{\gamma}^{(k)}$ are the same for all persons within class, they can vary over class. To illustrate, for a linear SPGM, Equation (1) would be as follows:

$$\eta_{ti} = \gamma_{00}^{(k)} + \gamma_{10}^{(k)} time_{ti} \text{ if } c_i = k \tag{3}$$

where intercept ($\gamma_{00}^{(k)}$) and slope ($\gamma_{10}^{(k)}$) growth coefficient values are the same for all persons in class *k* but differ across classes. We now turn to Equation (2). Equation (2) shows that the proportion of individuals in each class, $P(c_i = k)$, is estimated using a multinomial logistic regression where $\delta_0^{(k)}$ is the log odds of membership in class *k* versus the reference class. To identify this part of the model, $\delta_0^{(K)} = 0$, so the *K*th class is the reference class. (Time-invariant covariates could also be included by expanding Equation (2) to allow predictors of class membership.) The marginal Probability Density Function (PDF) of the SPGM is obtained by multiplying the class-specific conditional response distributions (either Bernoulli or normal) by their respective class membership probabilities and summing over classes:

$$f(y_{ti}|\mathbf{x}_{ti}) = \sum_{k=1}^{K} f(y_{ti}|\mathbf{x}_{ti}, c_i = k)P(c_i = k). \tag{4}$$

Hierarchical Linear and Generalized Linear Model

For HLM/HGLM, the *conditional response distribution* of the repeated measures is conditional on individuals' scores on predictor(s) with fixed effects ($\mathbf{x}_{ti}$), on individuals' continuously distributed deviations in growth, or random effects ($\mathbf{u}_i$), and on predictor(s) (e.g., time) with random effect(s) ($\mathbf{z}_{ti}$). Here, $\mathbf{z}_{ti}$ and $\mathbf{u}_i$ are $q \times 1$. For normal repeated measures $y_{ti}|\mathbf{x}_{ti}, \mathbf{z}_{ti}, \mathbf{u}_i \sim N(\mu_{ti}, \sigma^2)$ and for binary repeated measures $y_{ti}|\mathbf{x}_{ti}, \mathbf{z}_{ti}, \mathbf{u}_i \sim BERNOULLI(\mu_{ti})$. The *link functions* for HLM/HGLM are defined exactly as in SPGM. In HLM/HGLM, the *linear predictor* is as follows:

$$\eta_{ti} = \mathbf{x}'_{ti}\boldsymbol{\gamma} + \mathbf{z}'_{ti}\mathbf{u}_i. \tag{5}$$

To illustrate, for a linear HLM/HGLM with individual differences in intercepts and slopes, we would define $\mathbf{x}_{ti} = \mathbf{z}_{ti} = [1 \; time_{ti}]'$ and our linear predictor from Equation (5) would be

$$\eta_{ti} = \gamma_{00} + \gamma_{10} time_{ti} + u_{0i} + u_{1i} time_{ti} \tag{6}$$

where γ_{00} and γ_{10} are fixed effects representing the average intercept and linear slope, and u_{0i} and u_{1i} are random effects representing individual deviations in intercepts and linear slopes, respectively. Additionally, an aspect of the HLM/HGLM particularly relevant to this article is that unobserved random effects are assumed normally distributed, $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{T})$, where the (typically) unstructured $q \times q$ covariance matrix of the random effects is denoted $\mathbf{T}$. For example, with two random effects (random intercept and slope), we would have $\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$. The variability that is now accorded to the random effects in HGLM was accorded to class differences in the fixed effects in SPGM; hence the vectors $\mathbf{z}_{ti}$ and $\mathbf{u}_i$ did not appear in the SPGM. Thus, whereas in the SPGM the conditional response distribution was conditional on discretely distributed class membership ($c_i = k$), in HGLM it is conditional on the continuously distributed random effects ($\mathbf{u}_i$).

Finally, the marginal probability density function (PDF) for HLM/HGLM is obtained by integrating over (i.e., averaging over) the unobserved random effect(s):

$$f(y_{ti}|\mathbf{x}_{ti}, \mathbf{z}_{ti}) = \int f(y_{ti}|\mathbf{x}_{ti}, \mathbf{z}_{ti}, \mathbf{u}_i) f(\mathbf{u}_i) d\mathbf{u}_i. \tag{7}$$

Because the $\mathbf{u}_i$'s are unobserved, we can never be sure that $f(\mathbf{u}_i)$ is the normal PDF. As discussed later, worries about sensitivity of model estimates to misspecification of $f(\mathbf{u}_i)$ have motivated indirect interpretations of SPGM. The marginal PDF of the SPGM (Equation (4)) can be obtained by replacing the continuous density of the random effects $f(\mathbf{u}_i)$ in Equation (7) with a discrete class probability $P(c_i = k)$ and replacing the integral in Equation (7) with a sum across discrete classes. Thus, whereas HLM/HGLM arrives at a marginal distribution of the repeated measures by averaging (integrating over) continuous random effects, SPGM does so by averaging over discrete classes.

PRIOR RESEARCH ON THE ADEQUACY OF SPGM'S APPROXIMATION OF CONTINUOUS INDIVIDUAL DIFFERENCES

Having now defined the models of interest more formally, we return to the matter of what is known about the adequacy of SPGM's indirect approximation ability. To our knowledge, three prior studies have evaluated this question (Brame et al., 2006; Muthén & Asparouhov, 2008; Nagin, 2005) and did so in the following manner: Artificial data were generated from a model with continuously distributed random effects. An SPGM was fit to these data. SPGM estimates were then used to solve for general characteristics of the random effects distribution (mixed over groups) for comparison to the true, continuous distribution. Specifically—designating the model estimate for $P(c_i = k)$ as $\hat{\pi}^{(k)}$—masses and locations were used to obtain the overall mean(s) of the continuous random effect(s) in the generating population

$$\hat{\boldsymbol{\gamma}} = \sum_{k=1}^{K} \hat{\pi}^{(k)} \hat{\boldsymbol{\gamma}}^{(k)} \tag{8}$$

and to approximate (co)variance(s) of those continuous random effect(s) using between-class mean differences (e.g., Bauer, 2007; Vermunt & Van Dijk, 2001):

$$\hat{\mathbf{T}} = \sum_{k=1}^{K-1} \sum_{j=k+1}^{K} \hat{\pi}^{(k)} \hat{\pi}^{(j)} (\hat{\boldsymbol{\gamma}}^{(k)} - \hat{\boldsymbol{\gamma}}^{(j)})(\hat{\boldsymbol{\gamma}}^{(k)} - \hat{\boldsymbol{\gamma}}^{(j)})'. \tag{9}$$

These approximated moments were then compared with the true random effect means and (co)variances used to generate the data. Of course, the adequacy of SPGM's indirect approximation need not only be assessed for the lower order moments of a random effect distribution. Hence, in two studies, the cumulative distribution function (CDF) implied by the SPGM was also plotted against the true CDF to observe recovery of the distribution as a whole.

These studies indicated that SPGMs can adequately recover the mean and variance of a continuous random effect distribution using several classes (e.g., $K = 3$–6, with K selected by the Bayesian information criteria, BIC; Akaike information criteria, or a modified likelihood ratio test). For example, at $N =$

500, 2,500, and 10,000, Brame et al. (2006) reported $\leq$ 5% absolute relative bias (ARB), and at N = 2,000 Muthén and Asparouhov (2008) reported $\leq$ 11% ARB for the approximated mean and variance. Additionally, Brame et al. (2006) and Nagin (2005) visually depicted a close correspondence between a theoretical (true) and average sample-estimated SPGM CDF for a continuous random effect at N = 2,500 or N = 100,000, respectively.

Based on these results, Nagin and Tremblay (2005a, p. 882; 2005c, p. 84) concluded that "simulation evidence reported in Brame, Nagin, & Wasserman (2006) and Nagin (2005) suggests that relatively few points of support [groups/classes] are required to approximate reasonably complex continuous distributions of trajectories." However, these three prior studies considered quite low dimension random effect distributions: either one-dimensional (e.g., random intercept only; Brame et al., 2006; Nagin, 2005[2]) or two-dimensional (e.g., random intercept and linear slope; Muthén & Asparouhov, 2008). These three prior studies also used relatively large samples (N = 100,000, 10,000, 2,500, 2,000, or 500). Though both conditions may not apply to most psychology applications of SPGM, they are crucial to the adequacy of SPGM's indirect approximation as follows.

Random Effect Dimensionality

Although a row of mass points (e.g., class-specific intercept coefficients) suffices to approximate a one-dimensional random effect distribution—as depicted in Figure 2—a grid of mass points (e.g., class-specific intercept/slope coefficient coordinates) is needed to approximate a two-dimensional random effect distribution. After two dimensions, the discrete approximation becomes more difficult to visualize in these terms. But we can instead visualize the approximation in terms of what the best fitting trajectory classes would *actually* look like when they are approximating continuous individual differences in a sample that were generated along one dimension (random intercept; Figure 3 Panel A), versus two dimensions (random intercepts and linear slopes; Figure 3 Panel B), versus three dimensions (random intercepts, linear slopes, and quadratic slopes; Figure 3 Panel C).[3] We can see that class trajectories differ in level to approximate continuous variation in intercepts (i.e., classes appear stacked in Panel A), whereas classes differ in level and rate of change to approximate continuous variation in two dimensions (i.e., classes can appear crisscrossed in Panel B), and classes differ in level, instantaneous change, and acceleration/deceleration to approximate three dimensions (i.e., class patterns seem very qualitatively

[2]Nagin (2005) used two random effects that were correlated at 1.0, which is statistically equivalent to one random effect. Brame et al. (2006) used one random effect (a random intercept).

[3]The three samples used in Figure 3 Panels A–C had continuous repeated measures and were generated under HLM simulation conditions described later, which included quadratic fixed effects.

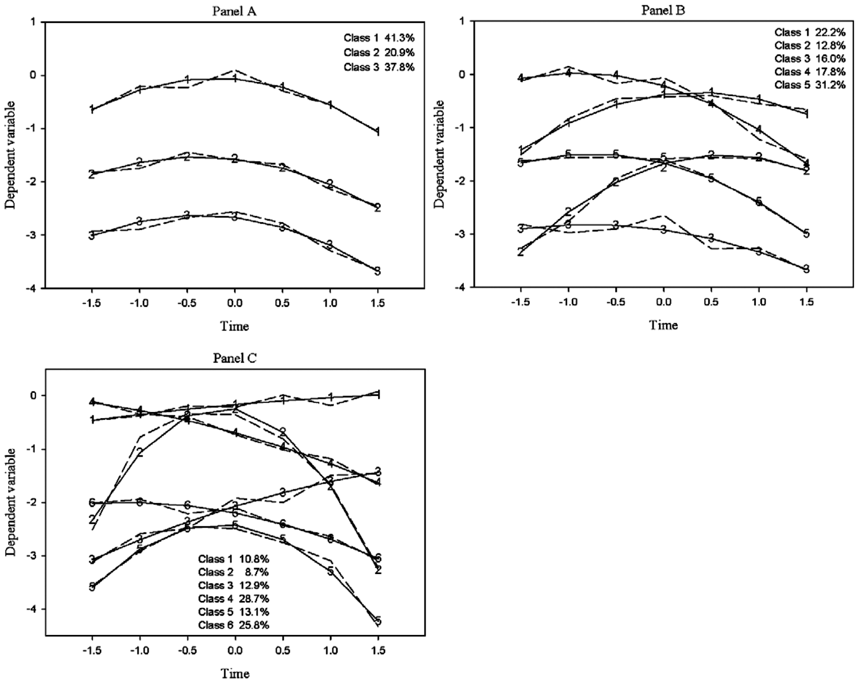

FIGURE 3 Illustration of semiparametric groups-based trajectory model indirect approximation of simulated unidimensional versus multidimensional continuous individual differences. *Note.* Panel A = SPGM fitted to a sample generated with a one-dimensional individual difference distribution (random intercepts); best Bayesian Information Criteria (BIC) $K = 3$. Panel B = SPGM fitted to a sample generated with a two-dimensional individual difference distribution (random intercepts and linear slopes); best BIC $K = 5$. Panel C = SPGM fitted to a sample generated with a three-dimensional individual difference distribution (random intercepts, linear slopes, quadratic slopes); best Bayesian Information Criteria $K = 6$. Solid lines = model estimated mean per class. Dashed lines = observed mean per class, calculated using modal class assignment. Proportions associated with each class are listed in the figure legend. Conditional response distribution was normal.

distinct in Panel C). It is important to note that the externalizing behavior SPGM applications in our review (comprising 22% of studies surveyed) contained at least one plot resembling Panels B or C; none exclusively contained classes resembling Panel A. This may suggest that classes are used to approximate high-dimension individual difference distributions in psychology practice. Further, though in applications researchers tend to intuitively perceive results like Panel A as more consistent with an indirect interpretation of classes but Panel C as more consistent with a direct interpretation of true population subgroups (e.g., Hadzi, 2009; Hill, White, Chung, Hawkins, & Catalano, 2000; Krueger, Markon, Patrick, & Iacono, 2005; Walton, Ormel, & Krueger, 2011), this is not necessary. As shown here, classes serving to indirectly approximate continua can give rise to plots like Panel A *or* C simply depending on the dimension of the underlying individual differences.

For higher dimension random effects, SPGM should need more classes to well approximate the surface. Potentially mitigating the effect of increasing random effect dimensionality is the correlation among random effects (Sterba, Mathiowetz, & Bauer, 2008). More tightly correlated random effects could begin to effectively function as a single dimension and require fewer classes to approximate. Yet no prior studies have compared the adequacy of SPGM's indirect approximation across different numbers of random effects or manipulated the correlation among random effects.

Sample Size

In our literature review of 100 psychology SPGM applications, sample sizes were considerably smaller than those used in prior evaluations of the adequacy of SPGM's indirect interpretation; 48% had $N < 500$ (the lowest sample size previously considered) and 77% had $N \leq 1{,}000$. Although the number of SPGM classes *needed* for an adequate indirect approximation may generally increase with the dimensionality of the random effect distribution, the number of SPGM classes actually extracted in practice (K) is effectively limited by N and model complexity. Specifically, K is typically selected by information criteria, such as BIC, which penalize for model complexity, especially at small N. Model complexity (loosely measured by the number of estimated parameters) is higher when class-varying quadratic and cubic terms are required.[4] Most (78%) of SPGM applications surveyed rely on BIC as the *sole* overall model selection index for choosing K. Although BIC is a consistent estimator of the number of classes (when the true distribution is a mixture; Leroux, 1992; Roeder & Wasserman, 1997), in small samples BIC tends to pick a K corresponding to a parsimonious, though not necessarily true, latent component structure (Nylund,

[4]A similar curse of dimensionality has been noted for related methods, such as nonparametric maximum likelihood estimation (NPMLE; Follmann & Lambert, 1989; Heckman & Singer, 1984; Laird, 1978) wherein 6–7 points of support may be needed to adequately approximate one random effect (e.g., Rabe-Hesketh, Pickles, & Skrondal, 2003) whereas 15 points of support may be needed to approximate two random effects (e.g., Schafer, 2001). But beyond two dimensions, "little is known about the performance of NPMLE for models with a large number of latent variables [random effects]" (Skrondal & Rabe-Hesketh, 2004, p. 183). However, in contrast to SPGM, the number of mass points K available for NPMLE is not as directly limited by N and model complexity because K is not chosen using model selection indices such as BIC (Lindsay, 1995; Skrondal & Rabe-Hesketh, 2004).

Asparouhov, & Muthén, 2007; Tofighi, & Enders, 2007).[5] Hence, enough mass points for an adequate indirect approximation of higher dimension individual differences may be inestimable, inefficient, or not optimally fitting at the modest *Ns* used in psychology. The previously studied combination of low random effect dimensionality and large *N* may constitute an ideal setting for the success of SPGM's indirect approximation.

PRIOR RESEARCH PERTAINING TO THE COMPARISON OF SPGM'S INDIRECT APPROXIMATION OF NONNORMAL RANDOM EFFECTS TO A PARAMETRICALLY MISSPECIFIED HLM

A key setting in which researchers find the indirect application of SPGM attractive relative to HLM is when the random effects are not multivariate normally distributed—as assumed by HLM (Nagin, 2005; Nagin & Tremblay, 1999, 2001, 2005a). Under this logic, errors due to the discrete approximation of the random effects distribution (with SPGM) would be tolerated in exchange for avoiding errors due to a parametric misspecification of the random effects distribution (with HLM). The question, therefore, is under what conditions will the error-of-approximation be larger, and under what conditions will the error-of-misspecification be larger?

Conditionally Normal Repeated Measures

For conditionally normal repeated measures, HLM's estimates (fixed effects and variance components) have been proven consistent regardless of the type/magnitude of random effect nonnormality, and simulations have supported their unbiasedness as well (Butler & Louis, 1992; McCulloch et al., 2008; Verbeke & Lessafre, 1997). Methods similar to indirectly applied SPGMs have also yielded unbiased fixed effects for conditionally normal repeated measures (Butler & Louis, 1992). Hence, for such repeated measures, together with nonnormal random effects, SPGM's error-of-approximation should only result in bias for variance components (increasing as the dimensionality of the random effects increases, especially at common sample sizes). In contrast, in this setting HLM should not incur error-of-misspecification for either fixed effects

[5] Unfortunately, overextracting classes in SPGM beyond the number selected as best fitting (by BIC) at a given *N* is not necessarily viable; this risks allowing a class proportion to approach zero or allowing parameters in two classes to approach the same values. Both situations can lead to singularities and estimation problems (McLachlan & Peel, 2000).

or variance components. The standard errors of HLM's estimates, and thus the validity of inferences, can be inconsistent when random effect normality is violated, but alternative standard error computations are robust to distributional violations (Verbeke & Lessafre, 1997). Nonetheless, we do not know how the efficiency of a misspecified HLM would compare with that of an SPGM indirect approximation.

Discrete Repeated Measures

Compared with conditionally normal repeated measures, prior research suggests a much more complex and nuanced picture of SPGM's and HGLM's relative performance under discrete repeated measures. For discrete repeated measures, fixed effects and variance components from both HGLMs and indirectly applied SPGMs can be biased when the distribution of the random effects is incorrect. Indeed, mean and covariance structures are dependent when modeling discrete outcomes such as binary and count, such that bias in the latter can affect estimates of the former (McCulloch et al., 2008). For discrete repeated measures, the degree of HGLM bias in fixed effects and variance components appears to depend on the type/magnitude of random effect nonnormality and magnitude of random effects. Specifically, for discrete repeated measures and small to medium random effects, HGLM has shown little/no fixed effects bias but produced biased variance components for some nonnormal random effect distributions (i.e., chi-square, log-normal, power) yet not others (bimodal-symmetric, uniform; e.g., Agresti, Caffo, & Ohman-Strickland, 2004; Butler & Louis, 1992; Litiere, Alonso, & Molenberghs, 2008; Neuhaus, Hauck, & Kalbfleisch, 1992; Rabe-Hesketh, Pickles, & Skrondal, 2003). When the random effects have very large variances, however, HGLMs have produced meaningfully biased estimates for both fixed effects and variance components regardless of the nonnormal distribution for random effects (Litiere et al., 2008).

Only one study compared the bias in a misspecified HGLM versus an indirectly applied SPGM. Using discrete repeated measures, Muthén and Asparouhov (2008) found HGLM's bias to be larger. However, they had severely nonnormal random effects—bimodal with modes 5–7 standard deviations apart. In contrast, when van den Oord et al. (2003) estimated random effect distributions for familiar psychological constructs (depression and delinquency) using two large national data sets, only modest nonnormality (skew range: −0.26 to 0; kurtosis range: −0.09 to 1.77) was found.[6] It is therefore important to clarify the relative performance of HGLM and SPGM under less severe nonnormality.

[6] Van den Oord et al. (2003) estimated random effect nonnormality by finding which of a family of Johnson Curves best fit the random effect distribution.

HYPOTHESES

Based on the prior review, the following hypotheses are posited about the performance of SPGM's indirect approximation and the performance of distributionally misspecified HLM/HGLMs.

Hypothesis 1

SPGM's indirect approximation of random effect variances (for conditionally normal or discrete repeated measures) and fixed effects (for discrete repeated measures) should suffer under more random effects, smaller *N*, and uncorrelated random effects. In particular, parameter bias will occur when too few classes are selected to achieve an adequate approximation; efficiency loss will occur when many classes are selected. Compared with HLM/HGLM, SPGM's performance should be relatively insensitive to nonnormality of random effects.

Hypothesis 2

HLM/HGLM's recovery of random effect (co)variances and fixed effects should show efficiency loss (for conditionally normal repeated measures) and bias (only for discrete repeated measures) under random effect nonnormality. Compared with SPGM, HLM/HGLM should be less sensitive to number of random effects, sample size, and random effect correlation.

Consistent with prior research, to test these hypotheses we generated data with continuous individual differences in change and fit SPGMs and HLM/HGLMs. To extend prior research, we manipulated random effect number, correlation, nonnormality (and type of nonnormality). This study also manipulated *N* and the conditional distribution of the repeated measures (discrete or continuous).

METHODS

Simulation Design

The simulation contained five design factors and a total of 90 cells. These design factors were not fully crossed for reasons described later. The first design factor was *N*: 250, 500, or 1,000. We chose these *N*'s because our review indicated that the majority of the *N*'s for SPGM applications in psychology fall within this range (only 16% of applications had $N < 250$; only 23% had $N > 1{,}000$). As mentioned later, however, we re-ran a subset of cells at $N = 10{,}000$ to further explicate some sample size effects. The second design factor was the number of random effects in the generating model: 1 (a random intercept only), 2 (random intercept and linear slopes), or 3 (random intercept, linear, and quadratic slopes). The third design factor was the type of distribution for the random intercept: *normal, skewed,* or *bimodal*. Other random effects, if present, were normally distributed. To further clarify distribution effects, however, we also re-ran a subset of cells with *all* random effects skewed or *all* random effects bimodal, as discussed later. The fourth design factor was the conditional response distribution of the repeated measures, which had two levels—Bernoulli (with probit link) or Normal (with identity link). The fifth design factor was the correlation among the random effects: none or moderate.

Population-Generating Models

Population-generating models were HLMs when the conditional response distribution was Normal and were HGLMs when it was Bernoulli, with the exception that the distributions of the random effects were not always normal. Five hundred samples were generated for the 90 simulation cells. Ruscio and Kaczetow's (2008) *R* program *GenData* was used to generate random effects data with the desired distributions and correlations, followed by SAS IML to generate seven repeated measures.

Parameters for the generating models were chosen with three objectives in mind. First, the linear predictor (i.e., $\eta_{ti} = \mathbf{x}'_{ti}\boldsymbol{\gamma} + \mathbf{z}'_{ti}\mathbf{u}_i$) was the *same* across binary and normal repeated measures. Also for normal repeated measures, the variance of the conditional response distribution was set to $\sigma^2 = 1$ so estimates would be on the same scale as the probit model estimates with binary repeated measures. These features allowed us to compare SPGM approximation adequacy across cells without confounding the effects of other design factors with the effects of the response distribution. Second, fixed effects were chosen to make the proportions of endorsed repeated measures similar to the pattern found in most binary SPGM applications on problem behavior in our review. In this common endorsement pattern, most (e.g., 43–86%) of persons follow flat, approximately zero trajectories (e.g., Bobo, Klepinger, & Dong, 2007; Falck, Wang, & Carlson, 2007; Jackson & Sher, 2008; B. Jones, Nagin, & Roeder, 2001; D. J. Jones et al., 2010; Lacourse, Nagin, Tremblay, Vitaro, & Claes, 2003; Sher, Gotham, & Watson, 2004). Hence, in our simulation, most persons (around 50%) never exhibit the behavior, but some persons show model-implied trajectories of sharp or slow onset and/or, in some cases, desistance at later ages. Third, variance components were chosen to render the intercept several times more variable than the linear slope (a 5:1 ratio is reported to be common by Muthén & Muthén, 2002). The linear variance was in turn larger than the quadratic, following empirical findings from polynomial growth models in Raudenbush and Bryk (2002) and Snijders and Bosker (1999).

The generating values fulfilling these three design objectives were as follows: $\boldsymbol{\gamma} = [-1.25 \quad -0.2 \quad -0.3]'$ and $\mathbf{x}_{ti} = [1 \quad time_{ti} \quad time_{ti}^2]'$ where $time_{1i} \ldots time_{7i}$ scores were $-1.5, -1, -.5, 0, .5, 1, 1.5$. $\mathbf{z}_{ti}$ was either $= [1]'$ or $[1 \quad time_{ti}]'$ or $[1 \quad time_{ti} \quad time_{ti}^2]'$ for 1, 2, or 3 random effects, respectively. For models with 1 random effect, $\tau_{00} = 1$. For 2 random effects, $\tau_{00} = 1$; $\tau_{11} = .15$. For 3 random effects, $\tau_{00} = 1$; $\tau_{11} = .15$; $\tau_{22} = .12$. In the *normal* random effect condition the intercept random effect was $N(0,1)$, in the *skewed* condition it was χ^2 (3), and in the *bimodal* condition it had a 3 standard deviation mean separation between modes. A nonnormal intercept random effect was subsequently transformed to have a mean of zero and the desired τ_{00}. The *skewed* condition (skew = 1.63, kurtosis = 4) mirrored the degree and type of random effect nonnormality detected in an application of HLM to antisocial behavior (Curran, 1997), which had skew = 1.16 and kurtosis = 4.99 when casewise Ordinary Least Squares intercepts were estimated and plotted as in Carrig, Wirth, and Curran (2004). The bimodal random effect distribution (skew = 0, kurtosis = −.96) might be more typical of constructs like vocabulary development (e.g., Bauer, Goldfield, & Reznick, 2002). Moderate random effect correlations were $\begin{bmatrix} 1 & -.6 \\ -.6 & 1 \end{bmatrix}$ for 2 random effects and $\begin{bmatrix} 1 & -.6 & -.6 \\ -.6 & 1 & .6 \\ -.6 & .6 & 1 \end{bmatrix}$ for 3 random effects.

Fitted Models

Model estimation was performed in Mplus 5.2 (Muthén & Muthén, 1998–2010). Data analysis was performed in SAS 9.2. Fitted models were either HLM/HGLM (distributionally misspecified for the nonnormal random effect conditions) or SPGM. When a SPGM was fitted, 300 sets of initial-stage random starting values were used, followed by 20 optimizations, in order to decrease the possibility of local maxima (Hipp & Bauer, 2006). Increasing numbers of classes were estimated starting with $K = 2$ and ending with any of the following problems: (a) nonconvergence, (b) a singularity or nonpositive definite covariance matrix, (c) parameter(s) fixed in order to avoid singularities, or (d) a class that became unstable and nearly collapsed into another class (i.e., a class proportion $< 1\%$). Maximum likelihood was used for SPGM and HLM along with numerical integration (adaptive rectangular quadrature; 15 integration points per dimension) for HGLMs.

Data Analysis

To evaluate Hypothesis 1, we assessed SPGM's recovery of (a) the first and second moments of the random effects distribution (as done by Muthén and

Asparouhov [2008] and Brame et al. [2006]), (b) the CDFs of the random effects distribution (as done by Brame et al. [2006]), and (c) the PDFs of the random effects distribution. Specifically, for several different values of *K,* we solved for population-level means and variances of the random effects distribution (i.e., γ_{00}, γ_{10}, γ_{20}, τ_{00}, τ_{11}, τ_{22}) from the SPGM output in replication *r* using Equations (7) and (8). This yielded indirectly approximated estimates $\hat{\gamma}_{00}$, $\hat{\gamma}_{10}$, $\hat{\gamma}_{20}$, $\hat{\tau}_{00}$, $\hat{\tau}_{11}$, $\hat{\tau}_{22}$ per replication *r.* Denote a generic estimate from this list as $\hat{\theta}_r$ (for replication *r*) and its population parameter as θ. At the best fitting *K,* according to the BIC, its absolute relative bias was $ARB = |(E(\hat{\theta}_r) - \theta)/\theta|$ and its mean square error was $MSE = E[(\hat{\theta}_r - \theta)^2]$ where the across-replications average was used in place of the expected value operator *E.* Best fitting *K* was chosen by the BIC because the BIC was used to select *K* in virtually all SPGM applications surveyed—usually exclusively. However, because key results are presented for *all* stably estimated solutions, rather than just for the best BIC solution, they are not contingent on the use of a particular model selection index. Additionally CDF and PDF plots give a holistic depiction of SPGM's approximation capabilities. For instance, we plotted the across-samples average of the empirical CDF (along with the intervals within which 90% of sample estimates fall) against the true CDF for the random effect distributions.

To evaluate Hypothesis 2, we calculated ARB and MSE of fixed effects and variance components estimates that were obtained directly from the output of possibly misspecified HLMs or HGLMs for replication *r.* Then we compared these ARB and MSE results between these possibly misspecified HLM/HGLMs and the indirectly approximated SPGMs.

RESULTS

To conserve space, tabled results only include uncorrelated random effect conditions. Corresponding tabled results for correlated random effect conditions are given in an Online Appendix. Figures include a combination of uncorrelated and correlated random effect conditions.

Continuous Repeated Measures

As anticipated by statistical theory, fixed effect estimates computed from continuous repeated measures were unbiased for both the distributionally misspecified HLM (here, ARB average = .0033; range = < .0001 to .0107) and the SPGM approximation (here, ARB average = .0033; range = < .0001 to .0124). Notably, fixed effects' MSE for distributionally misspecified HLMs (average = .0014; range = .0002 to .0056) was also equivalent to MSE for indirectly approximated

SPGMs (average = .0014; range = .0001 to .0056). Thus, for continuous (conditionally normal) repeated measures, we focus specifically on the variance component estimates and empirical CDF estimates.

SPGM

Convergence. For continuous repeated measures, few convergence problems were encountered when estimating SPGMs until well past the best BIC number of classes was extracted (typically 80–100% convergence until nine classes).

Random effect dimensionality. As an orientation to the continuous repeated measure SPGM results, first consider the PDF for one random effect with either a bimodal, skewed, or normal distribution at $N = 1{,}000$, in Figure 4. Comparing the across-samples average empirical PDF from SPGM versus the true population-generating PDF in Figure 4, we can see that for one random effect classes are essentially working as implied by the often-used heuristic sketch in Figure 2, regardless of distribution shape. However, consistent with Hypothesis 1, the approximation of the random effect covariance structure gets worse as the number of random effects increases, as shown in Table 1. Averaging across all cells, ARB of SPGM-approximated variance components was 2% for one random effect, 15% for two random effects, and 32% for three random effects. Figure 5 exemplifies these findings for a particular case, the skewed condition, hypothesized to be favorable to SPGM's variance approximation with respect to N and random effect correlation (i.e., $N = 1{,}000$, correlated random effects). Figure 5 plots SPGM's variance approximation at different K (star = best BIC K). The vertical bars are 90% estimate intervals at each K. The bold solid line connecting the vertical bar midpoints traces the average SPGM variance approximation across K. The horizontal thin solid line flanked by two dotted lines denotes the population-generating value and +/− 10% bias, respectively. Figure 5 shows that by three random effects, the approximation bias is sizable ($\geq$ 10% ARB), particularly for nondominant growth coefficients (i.e., coefficients with smaller variances)—even if we were to extract classes beyond the best BIC (starred) class. For nondominant growth factors, 90% estimate intervals do not even include the population-generating parameter at the best BIC number of classes.

Sample size. Table 1 also shows that when N increased, MSE for SPGM's variance approximation always decreased. Greater N often resulted in support for more classes (greater K), which typically translated into lower ARB for SPGM's variance approximation. When higher N supported more K, improvements to the approximation of entire random effect CDFs were noticeable. This result is

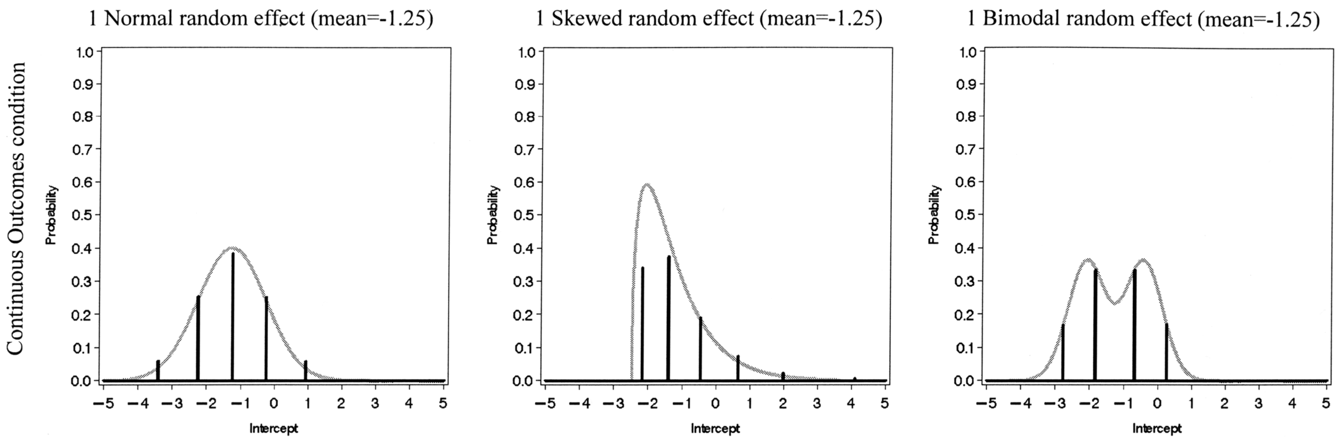

FIGURE 4 Semiparametric groups-based trajectory model (SPGM) approximation: Probability Density Function of a single dimension random effect distribution for normal repeated measures, $N = 1{,}000$, intercept mean $= -1.25$, variance $= 1$. *Note*. Grey line = theoretical (true) random effect distribution. Black bars = SPGM discrete approximation. The number of bars corresponds with the best Bayesian Information Criteria number of classes. The height of each bar is the across-samples average class probability for that class, and the location of each bar is the across-samples average intercept growth coefficient for that class.

TABLE 1
SPGM-Approximated Random Effect Variances: Continuous Outcomes

			Intercept Variance (1.0)			*Linear Variance (.15)*			*Quadratic Variance (.12)*		
Dist	*Samp*	*Best BIC*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*
1 RE											
N	250	4	.958	.014	.042						
	500	5	.991	.007	.009						
	1,000	5	.983	.004	.017						
S	250	4	.966	.029	.034						
	500	5	.981	.013	.019						
	1,000	6	.986	.007	.014						
B	250	4	.980	.009	.020						
	500	4	.985	.005	.015						
	1,000	4	.979	.003	.021						
2 REs, Uncorrelated											
N	250	5	.954	.005	.046	.084	.005	.438			
	500	7	.977	.006	.023	.118	.001	.211			
	1,000	8	.981	.004	.019	.124	.001	.174			
S	250	5	.961	.032	.039	.090	.005	.403			
	500	7	.985	.014	.015	.122	.001	.187			
	1,000	8	.982	.008	.018	.126	.001	.161			
B	250	5	.959	.010	.041	.100	.003	.332			
	500	6	.981	.006	.019	.116	.001	.225			
	1,000	7	.983	.003	.017	.126	.001	.159			
3 REs, Uncorrelated											
N	250	5	.833	.041	.167	.067	.008	.555	.030	.008	.749
	500	7	.878	.022	.122	.099	.003	.337	.042	.007	.651
	1,000	9	.906	.012	.094	.112	.002	.256	.061	.004	.493
S	250	5	.842	.051	.158	.069	.008	.539	.038	.007	.686
	500	7	.876	.028	.164	.101	.003	.633	.049	.008	.745
	1,000	9	.906	.016	.094	.114	.002	.242	.066	.003	.451
B	250	5	.866	.028	.134	.074	.007	.505	.039	.007	.671
	500	6	.867	.023	.133	.097	.003	.355	.039	.007	.673
	1,000	8	.901	.013	.099	.112	.002	.255	.060	.004	.500

Note. Dist = random effect distribution condition: (N = normal, S = skewed, or B = bimodal); RE = random effect; Samp = sample size; MSE = mean square error; ARB = absolute relative bias; BIC = Bayesian information criteria; SPGM = semiparametric groups-based trajectory model.

shown in Figure 6, which compares the true CDF versus across-samples average empirical CDF of a skewed or bimodal intercept in a correlated three random effect model at $N = 250$ (best K's 4 or 5) versus $N = 1{,}000$ (best K's = 6 or 7). To help visualize sampling variability around the across-samples average empirical CDF, 10 single-sample CDFs are also depicted at $N = 1{,}000$.

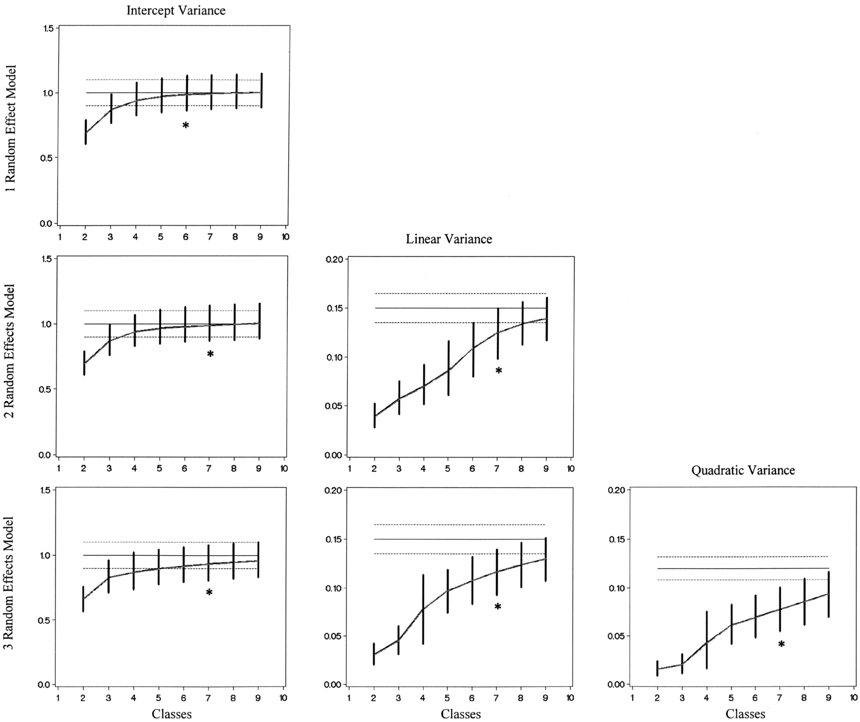

FIGURE 5 Semiparametric groups-based trajectory model approximation of random effect variances for continuous outcomes: $N = 1{,}000$, correlated, skewed random effect condition. *Note*. Horizontal thin flat solid line = parameter; horizontal thin flat dotted line = +/− 10% bias; length of bold vertical bar(s) = 90% estimate interval(s). Bold solid line connecting vertical bars = across-samples average. Star = # of classes preferred by Bayesian information criteria.

Correlation among random effects. Consistent with Hypothesis 1, one fewer class was typically best fitting for correlated (Online Appendix Table A) versus uncorrelated (Table 1) random effects. However, advantages in ARB and MSE for correlated random effects were only seen at three random effects (where ARB averaged 39% for uncorrelated vs. 26% for correlated random effects).

HLM

Convergence. No problems were encountered fitting HLMs to continuous repeated measures.

Nonnormality of random effects. For variance components, MSE was either the same or smaller for HLMs than SPGMs regardless of random effect

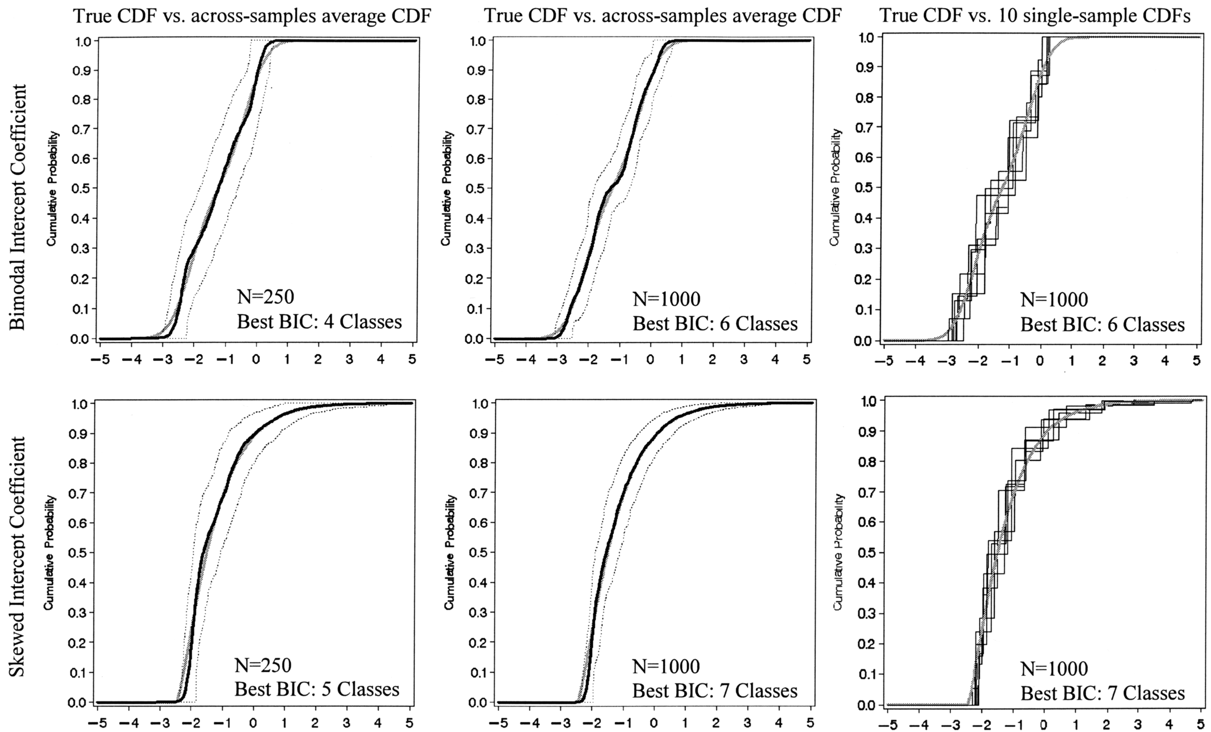

FIGURE 6 Semiparametric groups-based trajectory model approximation of the dominant (intercept) random effect Cumulative Distribution Function (CDF) for skewed or bimodal conditions with three correlated random effects and normal outcomes: $N = 250$ versus 1,000. *Note.* Thick grey line = theoretical (true) CDF. In columns 1 and 2, the thick black line = across-samples average empirical CDF; dotted lines = 90% estimate interval. In column 3, the 10 thin black lines = 10 single-sample empirical CDFs. BIC = Bayesian Information Criteria.

nonnormality (HLM average MSE = .0044; range MSE = .00015–.0280). For variance components, ARB was always trivial regardless of number of random effects for HLMs, consistent with statistical theory (average ARB = .0064; range ARB = .00003–.0275). Thus necessarily HLM's ARB was less than SPGM's ARB.

Binary Repeated Measures

SPGM

Convergence. For binary repeated measures, a "zero class" with a boundary constraint on the intercept coefficient was implemented to improve convergence above pilot rates.[7] There was 100% convergence for 2 classes, 80–100% for 3 classes, and 50–85% for 4-class models.

Random effect dimensionality. Comparing the across-samples average empirical PDF from SPGM versus the true population-generating PDF in Figure 7 ($N = 1{,}000$), classes no longer seem to be working as implied by the Figure 2 heuristic sketch. The spacing and heights of support points in Figure 7 appear much less optimal compared with the continuous condition in Figure 4. In particular, point masses (bars) in Figure 7 do not occupy the far left of the distributions but are overrepresented in the nearest point mass. This finding reflects SPGM's difficulty with empirically differentiating large negative intercept coefficient values that each imply almost all zero repeated measures. Consistent with Hypothesis 1, ARB of SPGM-approximated variance components on average increased with the number of random effects. As shown in Table 2, ARB increased from 22% to 42% to 60% for 1 to 2 to 3 random effects. More generally, Table 2 shows that SPGM's approximation is quite poor regardless of the number of random effects and regardless of the size of the variance of the random effect. Figure 8 illustrates this point for the skewed condition at $N = 1{,}000$ and correlated random effects; bias is always $\geq$ 10% at the best BIC number of classes.

[7]The simulation was piloted with the following free parameters in the fitted SPGM: one Level 1 error variance parameter (for continuous repeated measures only) and class-varying intercept, linear, and quadratic growth coefficients for each of K classes. However, for binary repeated measures, severe convergence problems were obtained during piloting when estimating more than two classes (even though almost always best BIC $K > 2$), and converged replications often had singularities and nonsensical, extreme growth coefficient values and/or standard errors. Diagnostic checks suggested that estimation problems often occurred when SPGM tried to reproduce the pattern of endorsement of individuals displaying mainly zeros over time. For these individuals there is little information from which to estimate the intercept (other than as a large negative value) or slopes (other than as nonpositive). In this light, we re-ran all binary cells with a designated flat class (linear and quadratic slopes fixed to zero) in which a boundary constraint of -3.5 was imposed on the intercept.

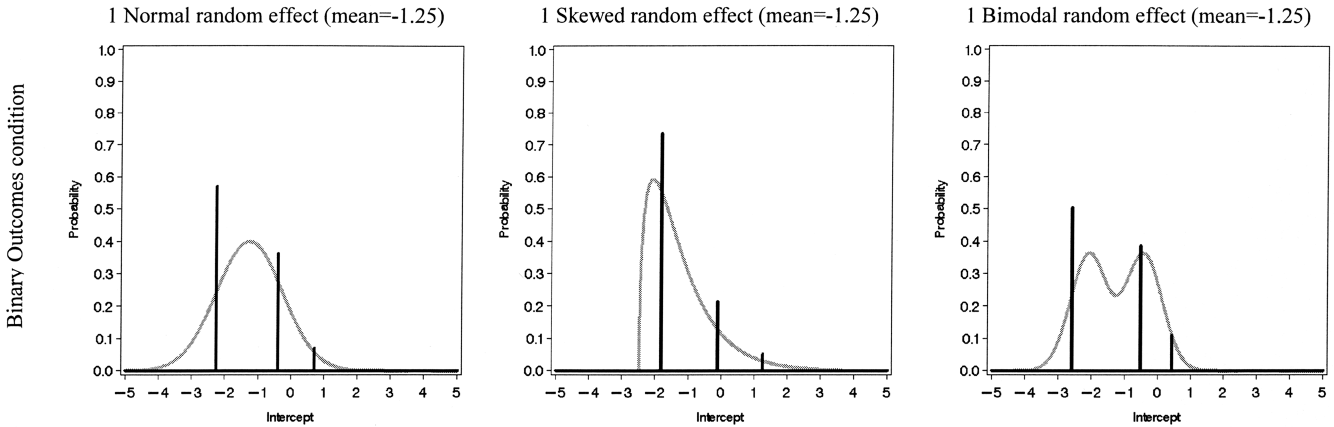

FIGURE 7 Semiparametric groups-based trajectory model (SPGM) approximation: Probability Density Function of a single dimension random effect distribution for binary repeated measures: $N = 1{,}000$, intercept mean $= -1.25$, variance $= 1$. *Note*. Grey line = theoretical (true) random effect distribution. Black bars = SPGM discrete approximation. The number of bars corresponds with the best Bayesian Information Criteria number of classes. The height of each bar is the across-samples average class probability for that class, and the location of each bar is the across-samples average intercept growth coefficient for that class.

TABLE 2
SPGM-Approximated Random Effect Variances: Binary Outcomes

Dist	*Samp*	*Best BIC*	*Intercept Variance (1.0)*			*Linear Variance (.15)*			*Quadratic Variance (.12)*		
			Estimate	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*
1 RE											
N	250	2	.526	.298	.474						
	500	3	1.136	.603	.136						
	1,000	3	1.098	.487	.098						
S	250	3	1.072	1.810	.072						
	500	3	.903	.300	.097						
	1,000	3	.838	.280	.162						
B	250	2	.777	.114	.223						
	500	3	1.329	.644	.329						
	1,000	3	1.428	.752	.428						
2 REs, Uncorrelated											
N	250	2	.545	.237	.455	.012	.019	.922			
	500	3	.934	.390	.066	.118	.650	.212			
	1,000	3	.918	.268	.082	.038	.114	.748			
S	250	3	1.056	.932	.056	.268	5.050	.785			
	500	3	.904	.262	.096	.044	.214	.708			
	1,000	3	.816	.119	.184	.008	.020	.950			
B	250	2	.690	.126	.310	.012	.020	.917			
	500	2	.685	.108	.315	.010	.020	.930			
	1,000	3	.949	.250	.051	.049	.013	.671			
3 REs, Uncorrelated											
N	250	2	.461	.315	.539	.011	.019	.925	.009	.013	.926
	500	3	.706	.615	.294	.072	.064	.519	.090	.222	.253
	1,000	3	.687	.232	.313	.023	.025	.848	.033	.080	.725
S	250	3	1.063	3.065	.063	.482	15.526	2.211	.243	1.023	1.491
	500	3	.797	.170	.203	.037	.090	.751	.021	.022	.828
	1,000	3	.728	.100	.272	.008	.020	.950	.006	.013	.951
B	250	2	.570	.219	.430	.014	.019	.909	.011	.012	.907
	500	2	.581	.188	.419	.011	.019	.930	.008	.013	.932
	1,000	3	.765	.530	.235	.037	.044	.752	.073	.709	.391

Note. Please refer to Table 1 notes.

Table 3 shows that, for binary repeated measures, the bias of the SPGM approximation also extends to the fixed effects portion of the model. As expected, the design conditions that resulted in greater variance component bias also resulted in greater fixed effect bias. In particular, fixed effect ARB was larger in the presence of more random effects (fixed effect ARB averaged 19% for one random effect, 27% for two random effects, and 35% for three random effects). Fixed effect ARB was worst for nondominant growth coefficients in models with multiple random effects.

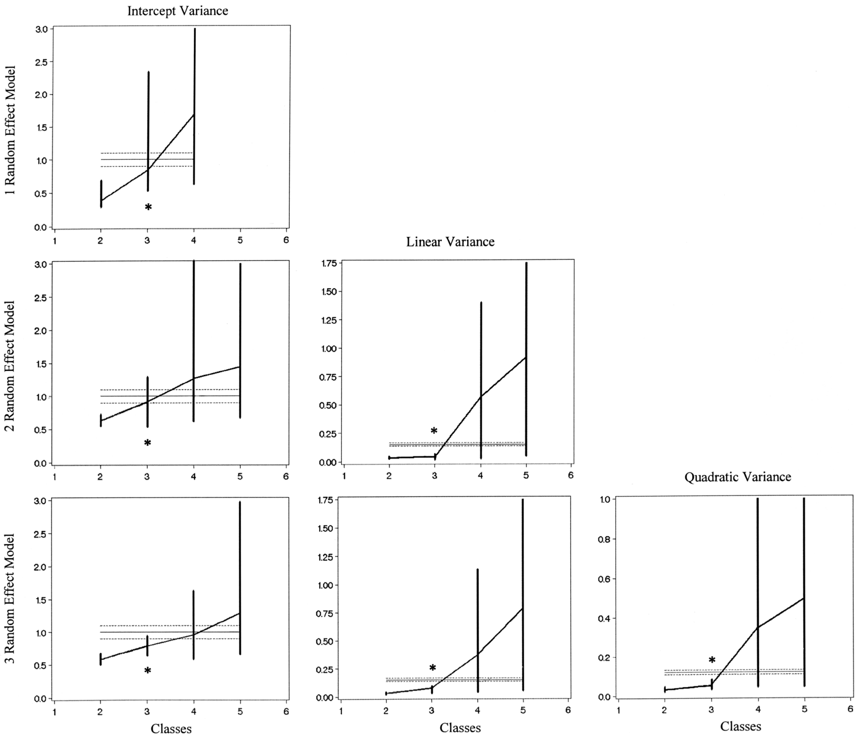

FIGURE 8 Semiparametric groups-based trajectory model approximation of random effect variances for binary outcomes: $N = 1{,}000$, correlated, skewed random effect condition. *Note.* Horizontal thin flat solid line = parameter; horizontal thin flat dotted line = +/− 10% bias; length of bold vertical bar(s) = 90% estimate interval(s). Black bold line connecting vertical bars = across-samples average. Star = # of classes preferred by the Bayesian information criteria. In some plots, the estimate intervals were truncated at ceiling values to ensure that 10% bias lines were visible.

Sample size. Increasing N often resulted in the selection of greater K as best fitting.[8] When larger N supported larger K, the additional classes typically translated into decreases in fixed effect and variance component ARB in Tables 2 and 3 and to improvements in recovery of the true generating CDFs. For instance,

[8]Overall, SPGM's ARB and MSE tended to be much higher in the binary outcomes condition than the normal outcomes condition. In case this might be due to binary outcomes requiring much higher N than continuous in order to achieve sufficient K for SPGM's approximation, we increased N to 10,000. We found the same pattern of results. We also calculated medians and trimmed means (i.e., most extreme 10% of sample estimates removed before averaging) to see if bias was meaningfully influenced by a few extreme sample estimates per cell. We again found the same pattern of results.

TABLE 3
SPGM-Approximated Fixed Effects: Binary Outcomes

			Intercept Mean (−1.25)			*Linear Mean (−.2)*			*Quadratic Mean (−.3)*		
Dist	*Samp*	*Best BIC*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*
1 RE											
N	250	2	−1.160	.027	.072	−.122	.015	.389	−.179	.028	.404
	500	3	−1.321	.079	.057	−.176	.078	.122	−.277	.099	.076
	1,000	3	−1.310	.068	.054	−.149	.005	.253	−.231	.009	.229
S	250	3	−1.312	.152	.050	−.147	.011	.264	−.364	.308	.213
	500	3	−1.284	.051	.027	−.171	.026	.147	−.267	.044	.112
	1,000	3	−1.263	.044	.010	−.170	.006	.149	−.255	.667	.149
B	250	2	−1.240	.014	.008	−.089	.018	.555	−.133	.035	.558
	500	3	−1.370	.085	.096	−.157	.053	.214	−.261	.075	.129
	1,000	3	−1.407	.093	.126	−.144	.006	.278	−.215	.011	.283
2 REs, Uncorrelated											
N	250	2	−1.168	.016	.065	−.091	.017	.546	−.109	.042	.638
	500	3	−1.247	.040	.002	−.136	.037	.321	−.179	.089	.403
	1,000	3	−1.252	.029	.002	−.125	.012	.374	−.155	.029	.484
S	250	3	−1.266	.052	.013	−.179	.073	.104	−.190	.061	.365
	500	3	−1.279	.033	.023	−.141	.012	.295	−.156	.027	.480
	1,000	3	−1.259	.015	.007	−.129	.007	.354	−.148	.025	.505
B	250	2	−1.188	.012	.050	−.085	.012	.575	−.109	.040	.635
	500	2	−1.191	.007	.047	−.074	.017	.632	−.096	.043	.681
	1,000	3	−1.257	.027	.006	−.118	.009	.411	−.144	.027	.520
3 REs, Uncorrelated											
N	250	2	−1.119	.026	.105	−.109	.014	.453	−.094	.048	.687
	500	3	−1.177	.048	.058	−.137	.023	.314	−.162	.068	.460
	1,000	3	−1.185	.022	.052	−.130	.012	.351	−.132	.052	.561
S	250	3	−1.308	.124	.047	−.167	.269	.425	−.171	.173	.742
	500	3	−1.261	.016	.009	−.131	.011	.343	−.092	.053	.693
	1,000	3	−1.245	.007	.004	−.129	.008	.355	−.086	.047	.713
B	250	2	−1.149	.021	.081	−.106	.016	.468	−.092	.049	.692
	500	2	−1.158	.013	.074	−.084	.016	.581	−.073	.054	.757
	1,000	3	−1.189	.037	.049	−.136	.015	.318	−.136	.079	.546

Note. Please refer to Table 1 notes.

in Figure 9, increasing N improved the average approximation of the intercept CDF when greater K was supported (e.g., skewed but not bimodal condition with three correlated random effects). However, this is not to suggest that more classes should be extracted at a given N even if not warranted by BIC. Rather, as K increased past the best BIC number at a given N, variance components went from underestimation to increasing overestimation in Figure 8, rather than getting closer to the population value. Furthermore, such "overextraction" of classes now involved extreme increases in sampling variability in Figure 8,

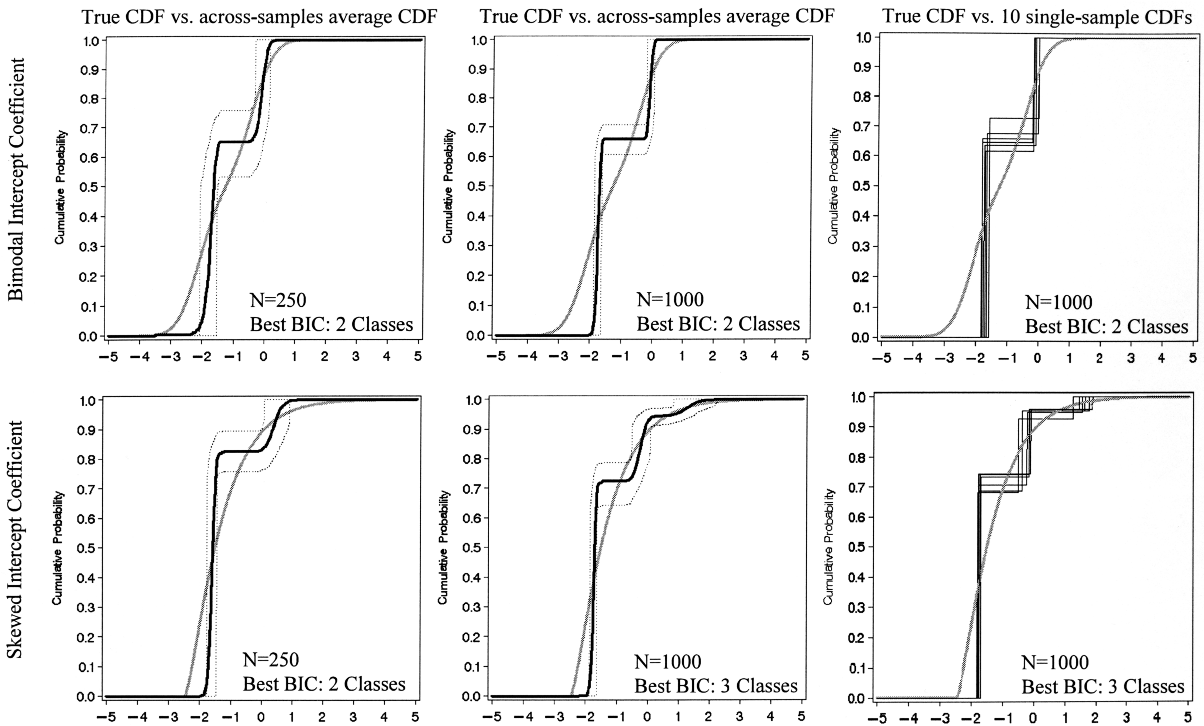

FIGURE 9 Semiparametric groups-based trajectory model approximation of the dominant (intercept) random effect Cumulative Distribution Function (CDF) for skewed or bimodal conditions with three correlated random effects and binary outcomes: $N = 250$ versus 1,000. *Note.* Thick grey line = theoretical (true) CDF. In columns 1 and 2, the thick black line = across-samples average empirical CDF; dotted lines = 90% estimate interval. In column 3, the 10 thin black lines = 10 single-sample empirical CDFs. BIC = Bayesian Information Criteria.

unlike for continuous outcomes. (However, MSE typically did decrease when *K* was held constant and *N* was increased—not shown in Tables 2 and 3).

Correlation among random effects. There was also some evidence for a decrease in bias associated with correlated (Online Appendix Tables B–C) versus uncorrelated (Tables 2 and 3) random effects, holding the number of random effects constant. Consistent with Hypothesis 1, average ARB for SPGM-approximated variance components for two uncorrelated versus correlated random effects was 47% versus 37% and for three uncorrelated versus correlated random effects was 70% versus 49%. Likewise, average SPGM-approximated fixed effect ARB for two uncorrelated versus correlated random effects was 32% versus 23% and for three uncorrelated versus correlated random effects was 39% versus 32%.

HGLM

Convergence. Few to no problems were encountered fitting HGLMs to binary outcomes.

Nonnormality of random effects. Tables 4 and 5 depict ARB and MSE when an HGLM is fit despite potential distributional assumption violations for random effects (see also Online Appendix Tables D–E for correlated random effect conditions). On average, ARB was 5% for fixed effects and 18% for random effect variances across alternative number of random effects, correlation among the random effects, and *N*. However, in line with Hypothesis 2, bias was affected by the random effect distribution. The normal and bimodal random effect conditions rarely showed sizable bias (average ARB = 1% and 3% for fixed effects, average ARB = 3% and 7% for random effect variances, respectively). The skewed condition often showed sizable bias (average ARB = 4% for fixed effects, average ARB = 40% for random effect variances). For the skewed condition, bias was most notable for the mean and variance of the intercept coefficient (the dominant and nonnormal coefficient). For this skewed intercept coefficient, HGLM's bias (in Tables 4 and 5) was greater than SPGM's (in Tables 2 and 3). For the other coefficients (which were nondominant and normal), HGLM's bias was less than SPGM's. Additionally, HGLM's MSE typically ranged from equivalent to better than SPGM but could be worse for the intercept coefficient's mean and variance in the skewed condition.

Generalizability Checks

All simulations are limited in external validity by their chosen conditions. Follow-up analyses ascertained whether our binary SPGM results were generaliz-

TABLE 4
Hierarchical Generalized Linear Model Random Effect Variances: Binary Outcomes

		Intercept Variance (1.0)			*Linear Variance (.15)*			*Quadratic Variance (.12)*		
Dist	*Samp*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*
1 RE										
N	250	1.007	.038	.007						
	500	1.007	.019	.007						
	1,000	1.000	.010	.000						
S	250	1.826	.844	.826						
	500	1.798	.730	.798						
	1,000	1.758	.611	.758						
B	250	.912	.033	.088						
	500	.925	.020	.075						
	1,000	.909	.016	.091						
2 RE, Uncorrelated										
N	250	1.001	.038	.001	.160	.007	.067			
	500	1.001	.021	.001	.151	.003	.008			
	1,000	1.002	.011	.002	.146	.001	.024			
S	250	1.748	.732	.748	.137	.007	.088			
	500	1.706	.572	.706	.136	.003	.093			
	1,000	1.693	.520	.693	.133	.002	.115			
B	250	.945	.039	.055	.166	.007	.105			
	500	.944	.018	.056	.164	.003	.096			
	1,000	.943	.012	.057	.160	.002	.067			
3 RE, Uncorrelated										
N	250	1.001	.043	.001	.154	.006	.029	.134	.010	.113
	500	1.008	.022	.008	.155	.004	.032	.126	.005	.050
	1,000	1.007	.013	.007	.149	.002	.007	.120	.002	.003
S	250	1.624	.533	.624	.151	.008	.008	.126	.010	.049
	500	1.620	.451	.620	.152	.004	.016	.127	.005	.056
	1,000	1.613	.409	.613	.148	.002	.015	.126	.002	.053
B	250	.967	.041	.033	.155	.007	.035	.127	.008	.059
	500	.958	.020	.042	.155	.003	.032	.127	.004	.054
	1,000	.963	.010	.037	.149	.002	.010	.121	.002	.011

Note. Please refer to Table 1 notes.

able beyond the kind of binary data typical of problem behavior applications and whether our HLM/HGLM results generalized to applications where all random effects were nonnormal, rather than just the random intercept.

SPGM generalizability check. For our SPGM generalizability check, we modified the endorsement pattern of our binary repeated measures to try to ease estimation issues. Our original binary data were not atypically sparse in the aggregate (as illustrated by the fact that HGLM had few convergence

TABLE 5
Hierarchical Generalized Linear Model Fixed Effects: Binary Outcomes

		Intercept Mean (−1.25)			*Linear Mean (−.2)*			*Quadratic Mean (−.3)*		
Dist	*Samp*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*	*Estimate*	*MSE*	*ARB*
1 RE										
N	250	−1.249	.011	.001	−.204	.003	.018	−.303	.003	.008
	500	−1.254	.005	.003	−.199	.001	.007	−.303	.002	.009
	1,000	−1.250	.003	.000	−.199	.001	.004	−.302	.001	.006
S	250	−1.565	.122	.252	−.201	.003	.006	−.303	.004	.010
	500	−1.561	.108	.249	−.200	.001	.002	−.302	.002	.006
	1,000	−1.549	.094	.239	−.201	.001	.007	−.304	.001	.013
B	250	−1.195	.013	.044	−.201	.003	.004	−.297	.003	.010
	500	−1.197	.008	.042	−.203	.001	.014	−.300	.002	.001
	1,000	−1.193	.006	.046	−.200	.001	.000	−.298	.001	.008
2 RE's, Uncorrelated										
N	250	−1.255	.012	.004	−.199	.003	.004	−.306	.004	.019
	500	−1.252	.006	.001	−.202	.002	.007	−.301	.002	.004
	1,000	−1.249	.003	.001	−.200	.001	.002	−.299	.001	.005
S	250	−1.533	.100	.227	−.202	.004	.008	−.291	.004	.029
	500	−1.529	.087	.223	−.203	.002	.013	−.287	.002	.042
	1,000	−1.522	.079	.217	−.200	.001	.001	−.288	.001	.040
B	250	−1.198	.014	.042	−.200	.003	.001	−.316	.004	.052
	500	−1.204	.007	.037	−.197	.002	.014	−.306	.002	.021
	1,000	−1.201	.005	.039	−.198	.001	.008	−.306	.001	.020
3 RE's, Uncorrelated										
N	250	−1.237	.010	.010	−.204	.004	.022	−.314	.012	.047
	500	−1.249	.006	.001	−.202	.002	.012	−.305	.005	.016
	1,000	−1.250	.003	.000	−.198	.001	.012	−.299	.003	.002
S	250	−1.527	.097	.222	−.209	.004	.047	−.280	.012	.067
	500	−1.518	.081	.214	−.207	.002	.033	−.279	.005	.070
	1,000	−1.510	.072	.208	−.204	.001	.022	−.282	.003	.061
B	250	−1.205	.013	.036	−.198	.004	.010	−.310	.010	.035
	500	−1.205	.008	.036	−.203	.002	.017	−.313	.006	.043
	1,000	−1.205	.005	.036	−.196	.001	.022	−.309	.003	.028

Note. Please refer to Table 1 notes.

problems, discussed previously), but the data became sparse when SPGM split the aggregate distributions into smaller classes. We made the binary data less sparse in the aggregate in a subset of cells (N = 1,000, uncorrelated random effects) by increasing the mean intercept from −1.25 to 0, implying that fewer (17%) of persons follow flat nonendorsement trajectories. When these new binary data were fit by SPGM, on average the points of support now appeared more optimally located for approximating the theoretical PDF of growth coefficients (Online Appendix Figure A) but still less so than for the continuous data. Fixed effects were now adequately recovered on average not only by HGLM but also by SPGM—although SPGM recovery worsened with 3 random effects in the generating model (see Online Appendix Table F). However, variance components were still on average better recovered with HGLM than SPGM. For 1, 2, and 3 random effects, respectively, SPGM ARB averaged 16%, 27%, and 28% whereas HGLM ARB averaged 4%, 10%, and 6%. Also, on average, HGLM was still more efficient than SPGM (average $MSE = .01$ vs. .20).

HLM/HGLM generalizability check. To check HLM/HGLM generalizability, we allowed all three random effects to be nonnormal (skewed or bimodal) at all *Ns* in the three uncorrelated random effect condition. ARB and MSE results were mostly comparable to our original simulation (see Online Appendix Table G). For example, for binary outcomes, HGLM's fixed effect ARB now averaged .097 for skewed and .024 for bimodal (vs. .105 and .029 in our original simulation). Also for binary outcomes, HGLM's variance component ARB now averaged .410 for skewed and .039 for bimodal (vs. .228 and .035 in our original simulation). For continuous outcomes, HLM's ARB now averaged .005 and MSE now averaged .003—again comparable to our original simulation.

SUMMARY

Though SPGM's indirect approximation abilities have previously been demonstrated for low-dimensional, very nonnormal individual difference distributions at relatively large *N,* this study investigated the generalizability of SPGM's indirect approximation abilities under a wider set of conditions that may be more typical of psychological research (e.g., relatively smaller sample sizes, higher dimensional random effect distributions, possibly modestly nonnormal random effects). We compared the error-of-approximation stemming from using SPGM as a discrete approximation tool with the error-of misspecification stemming from distributionally misspecified HLM/HGLMs.

Consistent with Hypothesis 1, more random effects, uncorrelated random effects, and smaller *N* (when smaller *N* translated into smaller best fitting *K*) on average increased bias for SPGM's indirect approximation. Such increases in bias were seen in the approximated variance components for continuous outcomes and in the approximated fixed effects and variance components for binary outcomes, as hypothesized. Also consistent with Hypothesis 1, SPGM was relatively insensitive to normality versus nonnormality of random effect distributions. Consistent with Hypothesis 2, random effect nonnormality increased bias for HGLMs in recovering the fixed effects and variance components for binary (but not continuous) outcomes. However, increased bias was mainly seen for the skewed rather than bimodal random effect condition. As hypothesized, HLM/HGLM were less sensitive to the number of random effects and correlation among random effects than SPGM.

Several additional findings were notable. First, for binary outcomes SPGM's strength was recovering means/variances of the dominant random effect (i.e., that with highest variance), whereas HGLM's was recovery of less dominant random effects—even when all random effects were made nonnormal in a follow-up generalizability study. Second, although extracting classes past best BIC improved SPGM's indirect approximation for continuous outcomes, it led to worse bias and extreme sampling variability for binary outcomes (Figure 5 vs. Figure 8). Third, compared with HGLM, binary SPGM led to strikingly more estimation problems—likely aggravated by high sparseness in some classes.

Finally, the results allowed us to weigh SPGM's error-of-approximation against HLM/HGLM's error-of-misspecification. For conditionally normal outcomes (used in 65% of SPGM applications in our review), HLM had equal or lower ARB and MSE than SPGM across cells. Thus, HLM's error-of misspecification $\leq$ SPGM's error-of-approximation even when random effects were modestly nonnormally distributed. Regardless of the number of classes selected, SPGM ARB was unacceptable ($>$ 10%) for some approximated variance components when there were two or three random effects. Hence, an approximation alternative to HLM seems not ultimately needed for conditionally normal repeated measures.

An approximation alternative to HGLM had greater potential to be useful for binary repeated measures; a distributionally misspecified HGLM can incur bias under this circumstance. However, for binary outcomes, HGLMs on average provided more accurate fixed effect and equivalent or better variance component estimates than SPGM's indirect approximation, even when it was not the true model. Further, HGLM estimates were typically as or more efficient than SPGM, except if the dominant growth coefficient was skewed. In sum, for binary outcomes, SPGM's indirect approximation was advantageous with one-dimensional skewed individual differences; HGLM's error-of-misspecification was more tolerable in most other circumstances.

EMPIRICAL ILLUSTRATIONS

Here we consider the insights our simulation results provide for practice, using empirical examples. Figure 1 depicts a $K = 6$ (best BIC K) cubic SPGM with a normal conditional response distribution applied to six repeated measures of externalizing behavior on $N = 250$ boys at ages 2, 3, 4.5, 6, 7, and 9 years from the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care (see NICHD Early Child Care Research Network,

TABLE 6
Externalizing Behavior Empirical Illustration: SPGM Indirect Approximation and HLM/HGLM Results

	Continuous Repeated Measures				*Binary Repeated Measures*			
	HLM		*SPGM-Approximation*		*HGLM*		*SPGM-Approximation*	
	M	*Variance*	*M*	*Variance*	*M*	*Variance*	*M*	*Variance*
Intercept	4.471*	9.464*	4.450	9.016	−4.597*	10.628*	−3.359	2.760
Linear	−0.443*	0.402*	−0.428	0.327	−0.678*	0.156	−0.448	0.0001
Quadratic	0.014	0.026*	0.018	0.009	0.037	0.028	0.058	0.001
Cubic	0.002	0.003*	−0.0002	0.002	0.028	—	0.011	—

Note. SPFM = semiparametric groups-based trajectory model; HLM = hierarchical linear model; HGLM = hierarchical generalized linear model. SPGM approximation used Equations (8) and (9). Time-specific residual variance estimates not shown. Significance level reported only for HLM/HGLM, where parameters were directly estimated. SPGM-approximated and HLM/HGLM estimated random effect covariances are provided in Online Appendix Table H.

$*p < .05$.

2004, for study design).[9] In Figure 1, the externalizing behavior measure was a total score of 14 externalizing Child Behavior Checklist (CBCL) items common to both the CBCL 2–3 and CBCL 4–18 (Achenbach, 1991, 1992). Researchers encountering results like Figure 1 often consider the possibility that classes may be indirectly approximating continuous individual differences (e.g., Feldman et al., 2009; Mazza et al., 2010; Murphy et al., 2009; Nagin, 2005; Segawa et al., 2005; Skardhamar, 2010; Xie et al., 2010). Understanding the implications of our simulation results requires first being able to visualize how this indirect approximation could manifest in a given plot. For Figure 1, a researcher can visualize that an underlying individual difference distribution would almost certainly be multidimensional because Figure 1 shows crisscrossing trajectories that do not look anything like Figure 3 Panel A. Hence, should an HLM be fitted, there will likely be meaningful variability in multiple random effects. And indeed the HLM estimates reported in Table 6 show statistically significant variability in intercepts and linear, quadratic, and cubic slopes of externalizing behavior.

[9] A subsetted sample size was chosen to mirror the simulation conditions; using the full sample can change the appearance and number of trajectory classes. Imposing a censored normal versus normal conditional response distribution yielded similar results. The empirical example SPGM and HLM with conditionally normal repeated measures used heterogeneous residual variances (σ_t^2). This example is for pedagogical purposes; related analyses are available elsewhere (e.g., for aggression; NICHD ECCRN, 2004).

Additionally, because a normal conditional response distribution is used, a researcher can infer from our simulation results that the indirect approximation of the *mean* externalizing trend by SPGM is accurate. But because the sample size is small, and the individual differences are multidimensional, a researcher can infer from our simulation results that the indirect approximation of the *variability* in externalizing behavior change by SPGM is inadequate even at the best BIC *K*. That is, although the large number of discrete classes, separation among the classes, and class differences in growth coefficients together attempt to capture the full range of individual variability in trajectories, they will underrepresent the true range of underlying variability. Indeed, we can gauge the degree of potential underestimation by applying the approximation formulas in Equations (8) and (9). Compared with the HLM estimates, these formulas indicate that SPGM accounts for 96%, 71%, 52%, and 62% of the variability in intercepts and linear, quadratic, and cubic slopes of externalizing behavior, respectively. Mirroring our simulation results, the SPGM will do best at approximating variation in the aspect of change that is most dominant (here, intercepts). We also have the assurance from our simulation results that the estimates of an HLM fitted to externalizing scores are accurate even if distributions of growth coefficients are actually nonnormal (a distributional misspecification).

To provide an example with binary data, we dichotomized the externalizing repeated measures to mimic clinical cutoff scores. Endorsing nine symptoms served as the clinical cutoff. Note this binary analysis is for illustrative purposes; dichotomizing sacrifices information and is not recommended. Figure 10 depicts a $K = 2$ (best BIC *K*) cubic SPGM with a Bernoulli conditional response distribution applied to these data. We again begin by visualizing how an indirect interpretation would manifest in Figure 10. These binary data are more impoverished than the continuous data such that the classes mainly differ in level—but somewhat in rate of change as well (though not enough to crisscross here). This implies that the underlying continuum of subject-specific trajectories of clinically significant externalizing is of lower dimension than in Figure 1. Correspondingly, when fitting an HGLM only three random effects are now estimable (rather than four) and the variance of only one of these is statistically significant (intercepts); see Table 6.

Further, a researcher can infer from our simulation results that, because the conditional response distribution is Bernoulli, SPGM's indirect approximation of both the mean trend and variance components can be unacceptably biased—particularly for capturing rates of change in clinically significant externalizing. Though in an empirical application we do not know true population values, applying Equations (8) and (9) suggests that SPGM's indirect approximation is capturing $< 1\%$ of the variation in linear slopes and 2% of the variation in quadratic slopes compared with the HGLM results. Further, the SPGM approx-

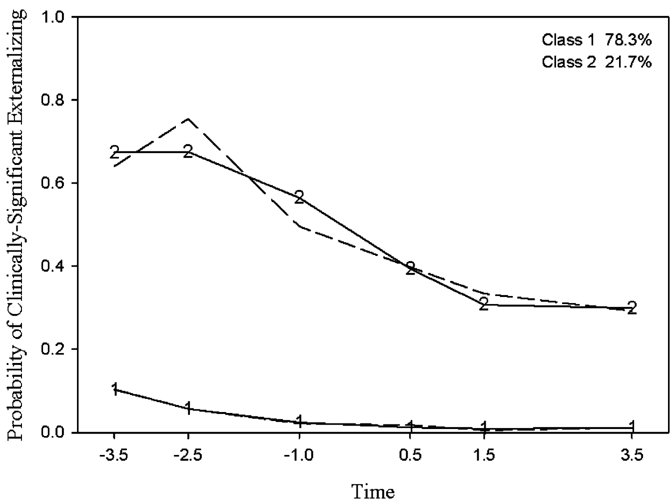

FIGURE 10 Semiparametric groups-based trajectory model of externalizing behavior clinical cutoff scores: Two-class empirical example. *Note.* Time = (age in years—5.5). Solid lines = model estimated mean per class. Dashed lines = observed mean per class, calculated using modal class assignment. Proportions associated with each class are listed in the figure legend. Illustrative class labels: 1 = "high-declining"; 2 = "low-stable." Conditional response distribution assumed Bernoulli. Details are given in the section of this article titled Empirical Illustrations.

imated average rates of change (fixed effects) are also under- or overestimated compared with the HGLM estimates in Table 6.

This suggests that the impoverished binary data is not able to support enough classes with class-varying growth coefficients to recover all change variation in the data. If the underlying individual differences distributions are skewed, a researcher can infer from our simulation results that HGLM will also be inaccurate—particularly for the intercept estimates. For instance, in Table 6 intercept dispersion may be overestimated by the HGLM. Stepping back, a researcher contemplating a semiparametric versus parametric model may consider where (e.g., intercepts or slopes) the theory necessitates most accurate recovery or may generally consider which model has best recovery on average (see Summary).

RECOMMENDATIONS FOR PRACTICE

1. *Provide more concrete details on an indirect interpretation.* Rather than just citing an indirect function of classes in the abstract, researchers can explain and concretely visualize how this approximation could manifest for different random

effect dimensionality. Previously, researchers intuitively tended to connect plots resembling Figure 3 Panel A with indirectly approximated *quantitative* individual differences and Panel C with directly interpretable *qualitative* individual differences. We emphasized how both plots can be consistent with an indirect approximation—just of different numbers of random effects. Recent literature reviews of dozens of externalizing behavior SPGM applications have failed to replicate basic descriptive aspects of trajectory groups (e.g., Fontaine, Carbonneau, Vitaro, Barker, & Tremblay, 2009; Horn, 2000; Skardhamar, 2009; van Dulmen, Goncy, Vest, & Flannery, 2009), generating discussion as to why this is so. For instance, in Fontaine et al.'s (2009) review, 5% of studies had over 5 classes, 29% had 5 classes, 28% had 4, 28% had 3, and 10% had 2. The proportions and shapes of these classes also varied widely (e.g., chronic (4%)—escalators (12%)—desistors (35%)—late-onsetters (17%)—nonoffenders (32%) vs. high-rising (35%)—low (65%) vs. high decreasers (4%)—low-decreasers (15%)—near-zero (81%)). From an indirect perspective on classes (Nagin & Tremblay, 2005a) this finding is not surprising;[10] for instance, it is known that altering the number of timepoints or using coarsely categorized repeated measures can affect the number of random effects supported by HLM/HGLM (Fitzmaurice, Laird, & Ware, 2011; Hedeker & Gibbons, 2006). In turn, differing extents of continuous interindividual variability from one data set to another would require different number(s) of classes, patterns of class-varying growth coefficients, and proportions of class membership for their indirect approximation.

2. *Consider when an indirect approximation will be adequate.* Researchers should not just cite an indirect interpretation of classes without giving consideration to whether their data conditions will yield an adequate approximation of continuous individual differences. Generally, this approximation is worse, and often inadequate, with lower N, multidimensional individual differences, and/or binary repeated measures. For continuous repeated measures, researchers can quantify and report the adequacy of approximation in their specific data set by applying Equations (8) and (9) to their SPGM results and comparing these numbers to HLM results. Researchers should keep in mind that this is a sample-level comparison; though HLM estimates will be consistent under distributional violation, they, as well as SPGM estimates, will still naturally be subject to sampling variability. Class-varying fixed coefficients in the SPGM could be allowed fixed and random effects in the corresponding HLM. Example SAS code is provided in the Online Appendix to implement these computations. If the approximation is poor, a *direct* interpretation of classes—with its added strict assumption that classes correspond with true population subgroups (Bauer, 2007;

[10]Others have sought to explain these discrepant findings more from a direct perspective (e.g., Fontaine et al., 2009; Van Dulmen et al., 2009; see also Eggleston et al., 2004; Jackson & Sher, 2008).

Sterba & Bauer, 2010)—would still be available for researchers fitting an SPGM. If the approximation is good, *both* direct and indirect interpretations could provide adequate explanations from a statistical perspective. (For binary repeated measures, such a comparison would be less definitive because HGLM results could be biased under distributional violation, particularly for the intercept coefficient; researchers could instead simply gauge approximation adequacy by consulting the most relevant of our simulation cells.)

Researchers should also be aware that some common practices used to simplify *direct* interpretability of classes can actually cause the indirect approximation to suffer. Researchers often choose to retain *fewer* classes than that recommended by statistical fit indices on grounds that more classes do not aid direct substantive interpretability. For instance, it is more difficult to assign unique verbal descriptor labels to multiple classes of similar functional form and similar level. So when adding a class splits an existing class into slightly separated classes of similar functional form, researchers may prefer the $K-1$ solution even if fit indices prefer the K class solution (e.g., Brame, Nagin, & Tremblay, 2001; Gross, Shaw, Burwell, & Nagin, 2009; Otten, Wanner, Vitaro, & Engels, 2008; Petitclerc, Boivin, Dionne, Zoccolillo, & Tremblay, 2009; van der Vorst, Vermulst, Meeus, Dekovic, & Engels, 2009).[11] Researchers need to keep in mind that although this practice may make a direct interpretation easier and perhaps more substantively compelling, it makes an indirect interpretation of classes more unrealistic. For instance, in our simulated Figure 3 Panel C example, extracting one fewer class than best BIC results in classes #5 and #6 collapsing into one. Because the merged class had a shape similar to the originals, it could be given a similar label ("moderate level; concave down"); thus from a direct interpretation vantage point there may be little substantive loss. But from an indirect interpretation vantage point, there is a cost: individual variability in growth coefficients would be underestimated (approximation adequacy dropped by 17% for linear slope variance and 15% for quadratic slope variance when K was 1 less than best BIC). When researchers choose reduced K to aid direct

[11]For instance, Petitclerc et al. (2009) explain that "with four groups and more, a small, high stable group was consistently found, and adding groups resulted in splitting lower level groups. Therefore, the four-group solution was retained" (p. 1479); likewise Gross et al. (2009) write, "Despite improved BIC scores, both the five and six group models resulted in subdividing already modest size groups with higher levels of maternal depressive symptoms into smaller groups that were not substantively different from one another; thus, the four group model emerged as the best fitting and most parsimonious model" (p. 147). Beyers & Seiffge-Krenke (2007) explain that "if a solution with K classes emerges in which certain classes are merely slight variations on a common theme and, hence, do not have differential substantive meaning, the more parsimonious solution with $K-1$ classes is chosen" (p. 563). Brame et al. (2001) similarly state, "For the adolescent aggression data a six-group model was found to best fit the data. However, here we describe the four-group model because the results from this more parsimonious solution are qualitatively similar" (p. 506).

substantive interpretability, but still desire an indirect interpretation of classes as a fallback explanation, they should mention the potential disadvantages of this practice. Moreover, researchers can quantify the drop in indirect approximation ability for specific aspects of change by applying Equations (8) and (9) to SPGM results before and after leaving out a class.

3. *Consider when an indirect approximation will be preferable.* This study identified some intriguing trade-offs. On one hand, HGLM suffers under the specification error of random effect nonnormality. On the other hand, SPGM's approximation suffers specifically under high dimensional random effect distributions and lower N and generally under binary data (particularly when it can become sparse when divided into classes). For the kind of data examined here—which mirrors developmental psychopathology applications—HLM and HGLM performed adequately under a wider variety of circumstances than did SPGM. Hence, it is not always the case that "if a suitable continuous distribution is not known or tractable, the group-based, semiparametric approach is an attractive alternative" (Nagin & Tremblay, 2001, p. 29).

In practice, certain diagnostics can be employed to aid researchers in deciding whether their data are more susceptible to approximation error (and hence more conducive to HGLM) or more susceptible to specification errors discussed here (and hence more conducive to SPGM). For example, plots of ordinary least squares (OLS) individual trajectory estimates could perhaps suggest whether individuals vary in linear, quadratic, and/or cubic slopes in addition to intercepts—implying a high-dimension random effect distribution less conducive to SPGM (see Singer & Willet, 2002, for interpretation considerations involving degree of sampling variability). Also, by-timepoint frequencies of binary repeated measures may indicate a risk for sparseness upon subdivision into classes (less conducive to SPGM). Additionally, plots of model-implied (Empirical Bayes) random effect distributions could give some indication of normality violations (less conducive to HGLM). However, these predicted scores are shrunken toward a normal prior distribution during estimation, particularly when Level 1 residual variability is large (Verbeke & Lessafre, 1996). Alternative newer diagnostics for random effect normality are reviewed by Huang (2011).

CONCLUSIONS

SPGM's indirect approximation is often cited in methodological and applied work. Care must be taken when motivating its use based on an unqualified ability to approximate continuous possibly nonnormal individual differences; this approximation will sometimes not be adequate. Furthermore, this approximation may often not be necessary due to the robustness of HLM and even HGLM to moderate violations of distributional assumptions for random effects. Future research could investigate the accuracy of alternative approximation methods (Galindo-Garre & Vermunt, 2006; Magidson & Vermunt, 2001; Zhang & Davidian, 2001) in comparison to SPGM.

REFERENCES

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

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist 2–3 and 1992 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Agresti, A., Caffo, B., & Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, *47,* 639–653.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research, 42,* 757–786.

Bauer, D. J., Goldfield, B. A., & Reznick, J. S. (2002). Alternative approaches to analyzing individual differences in the rate of early vocabulary development. *Applied Psycholinguistics, 23,* 313–335.

Beyers, W., & Seiffge-Krenke, I. (2007). Are friends and romantic partners the "best medicine"? How the quality of other close relations mediates the impact of changing family relationships on adjustment. *International Journal of Behavioral Development, 31,* 559–568.

Bobo, J. K., Klepinger, D. H., & Dong, F. B. (2007). Identifying social drinkers likely to consume alcohol during pregnancy: Findings from a prospective cohort study. *Psychological Reports, 101,* 857–870.

Brame, R., Nagin, D. S., & Tremblay, R. E. (2001). Developmental trajectories of physical aggression from school entry to late adolescence. *Journal of Child Psychology and Psychiatry, 42,* 503–512.

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

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

Butler, S. M., & Louis, T. A. (1992). Random effects models with non-parametric priors. *Statistics in Medicine, 11,* 1981–2000.

Carrig, M., Wirth, R. J., & Curran, P. J. (2004). A SAS Macro for estimating and visualizing individual growth curves. *Structural Equation Modeling, 11,* 132–149.

Curran, P. J. (1997, April). *Comparing three modern approaches to longitudinal data analysis: An examination of a single developmental sample.* Symposium presented at the biennial meeting of the Society for Research in Child Development, Washington, DC.

Delucchi, K. L., Matzger, H., & Weisner, C. (2004). Dependent and problem drinking over 5 years: A latent class analysis. *Drug and Alcohol Dependence, 74,* 235–244.

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.

Falck, R. S., Wang, J., & Carlson, R. G. (2007). Crack cocaine trajectories among users in a midwestern American city. *Addiction, 102,* 1421–1431.

Feldman, B. J., Masyn, K. E., & Conger, R. D. (2009). New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental Psychology, 45,* 652–676.

Fitzmaurice, G., Laird, N., & Ware, J. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, NJ: Wiley.

Follmann, D. A., & Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association, 84,* 295–300.

Fontaine, N., Carbonneau, R., Vitaro, F., Barker, E. D., & Tremblay, R. E. (2009). Research review: A critical review of studies on the developmental trajectories of antisocial behavior in females. *Journal of Child Psychology and Psychiatry, 50,* 363–385.

Galindo-Garre, F., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika, 33*, 43–59.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.

Gross, H. E., Shaw, D. S., Burwell, R. A., & Nagin, D. S. (2009). Transactional processes in child disruptive behavior and maternal depression: A longitudinal study from early childhood to adolescence. *Development and Psychopathology, 21,* 139–156.

Hadzi, D. (2009). Finding patterns and groupings I: Introduction to latent class analysis. *Acta Neuropsychiatrica, 21,* 312–313.

Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica, 52,* 271–320.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York, NY: Wiley.

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

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

Horn, J. L. (2000). Comments on integrating person-centered and variable-centered research on problems associated with the use of alcohol. *Alcoholism: Clinical and Experimental Research, 24,* 924–930.

Huang, X. (2011). Detecting random-effects model misspecification via coarsened data. *Computational Statistics and Data Analysis, 55,* 703–714.

Jackson, K. M., & Sher, K. J. (2008). Comparison of longitudinal phenotypes based on alternate heavy drinking cut scores: A systematic comparison of trajectory approaches III. *Psychology of Addictive Behaviors, 22,* 198–209.

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

Jones, D. J., Runyan, D. K., Lewis, T., Litrownik, A., Black, M., Wiley, T., . . . Nagin, D. S. (2010). Trajectories of childhood sexual abuse and early adolescent HIV/AIDS risk behaviors: The role of other maltreatment, witnessed violence, and child gender. *Journal of Clinical Child & Adolescent Psychology, 39,* 667–680.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology, 24,* 1–31.

Krueger, R., Markon, K., Patrick, C., & Iacono, W. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-IV. *Journal of Abnormal Psychology, 114,* 537–550.

Lacourse, E., Nagin, D., Tremblay, R. E., Vitaro, F., & Claes, M. (2003). Developmental trajectories of boys' delinquent group membership and facilitation of violent behaviors during adolescence. *Development and Psychopathology, 15,* 183–197.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73,* 805–811.

Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics, 38,* 963–974.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics, 20,* 1350–1360.

Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications.* Volume 5 of NSF-CBMS Regional Conference Series in Probability and Statistics. Hayward, CA: Institute of Mathematical Statistics.

Litiere, S., Alonso, A., & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine, 27,* 3125–3144.

Losoya, S., Knight, G., Chassin, L., Little, M., Vargas-Chanes, O., Mauricio, A., & Piquero, A. (2008). Trajectories of acculturation and enculturation in relation to heavy episodic drinking and marijuana use in a sample of Mexican American serious juvenile offenders. *Journal of Drug Issues, 1,* 171–198.

Louvet, B., Gaudreau, P., Menaut, A., Genty, J., & Deneuve, P. (2007). Longitudinal patterns of stability and change in coping across three competitions: A latent class growth analysis. *Journal of Sport & Exercise Psychology, 29,* 100–117.

Maggi, S., Hertzman, C., & Vaillancourt, T. (2007). Changes in smoking behaviors from late childhood to adolescence: Insights from the Canadian national longitudinal survey of children and youth. *Health Psychology, 26,* 232–240.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology, 31,* 223–264.

Mazza, J. J., Fleming, C. B., Abbott, R. D., Haggerty, K. P., & Catalano, R. (2010). Identifying trajectories of adolescents' depressive phenomena: An examination of early risk factors. *Journal of Youth and Adolescence, 39*, 579–593.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models.* Hoboken, NJ: Wiley.

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

Murphy, D. A., Brecht, M. L., Herbeck, D. M., & Huang, D. (2009). Trajectories of HIV risk behavior from age 15 to 25 in the National Longitudinal Survey of Youth sample. *Journal of Youth Adolescence, 38,* 1226–1239.

Muthén, B. O. (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.

Muthén, B. O. (2008). Latent variable hybrids: Overview of new and old models. In G. Hancock & K. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age.

Muthén, B., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In F. G. Davidian, M. G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus users' guide* (5th ed.). Los Angeles, CA: Author.

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

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

Nagin, D. S. (2004). Response to "Methodological sensitivities to latent class analysis of long-term criminal trajectories." *Journal of Quantitative Criminology, 20,* 27–35.

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. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6,* 18–34.
Nagin, D. S., & Tremblay, R. E. (2005a). Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology, 43,* 873–904.
Nagin, D. S., & Tremblay, R. E. (2005b). From seduction to passion: A response to Sampson and Laub. *Criminology, 43,* 915–918.
Nagin, D. S., & Tremblay, R. E. (2005c). What has been learned from group-based trajectory modeling? Examples from physical aggression and other problem behaviors. *Annals of the American Academy of Political and Social Science, 602,* 82–117.
Nagin, D. S., & Tremblay, R. E. (2005d). Further reflections on modeling and analyzing developmental trajectories: A response to Raudenbush and Maughan. *Annals of the American Academy of Political and Social Science, 602,* 145–154.
Nash, J. K., & Kim, J. S. (2007). Patterns of change over time in beliefs legitimizing aggression in adolescents and young adults: Risk trajectories and their relationship with serious aggression. *Social Work Research, 31,* 231–240.
Neuhaus, J. M., Hauck, W. W., & Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika, 79,* 755–762.
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*.
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, 14,* 535–569.
Obradovic, J., Burt, K. B., & Masten, A. S. (2006). Pathways of adaptation from adolescence to young adulthood antecedents and correlates. *Annals of the New York Academy of Sciences, 1094,* 340–344.
Otten, R., Wanner, B., Vitaro, F., & Engels, R. (2008). Own and friends' smoking attitudes and social preference as early predictors of adolescent smoking. *Journal of Clinical Child & Adolescent Psychology, 37,* 808–819.
Petitclerc, A., Boivin, M., Dionne, G., Zoccolillo, M., & Tremblay, R. (2009). Disregard for rules: The early development and predictors of a specific dimension of disruptive behavior disorders. *Journal of Child Psychology and Psychiatry, 50,* 1477–1484.
Piquero, A. R., Fagan, J., Mulvey, E. P., Steinberg, L., & Odgers, C. (2005). Developmental trajectories of legal socialization among serious adolescent offenders. *Journal of Criminal Law and Criminology, 96,* 267–298.
Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling, 3,* 215–232.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
Rodriguez-Zas, S., Southey, B., Whitfield, C., & Robinson, G. (2006). Semiparametric approach to characterize unique gene expression trajectories over time. *BMC Genomics, 7,* 233–240.
Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association, 92,* 894–902.
Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research, 43,* 355–381.
Sampson, R. J., & Laub, J. H. (2003). Life course desisters? Trajectories of crime among delinquent boys followed to age 70. *Criminology, 41,* 555–592.
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.
Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics, 57,* 53–61.
Segawa, E., Ngwe, J. E., Li, Y., Flay, B., & Coinvestigators. (2005). Evaluation of the effects of the Aban Aya Youth Project in reducing violence among African American adolescent males using latent class growth mixture modeling techniques. *Evaluation Review, 29,* 128–148.
Sher, K. J., Gotham, H. J., & Watson, A. L. (2004). Trajectories of dynamic predictors of disorder: Their meanings and implications. *Development and Psychopathology, 16,* 825–856.
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.
Skardhamar, T. (2009). Reconsidering the theory on adolescent-limited and life-course persistent antisocial behavior. *British Journal of Criminology, 49,* 863–878.
Skardhamar, T. (2010). Distinguishing facts and artifacts in group-based modeling. *Criminology, 48,* 295–320.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.
Sterba, S. K., & Bauer, D. J. (2010). Matching method with theory in person-oriented developmental psychopathology research. *Development & Psychopathology, 22,* 239–254.
Sterba, S. K., Mathiowetz, R. E., & Bauer, D. J. (2008). Abstract: Adequacy of semiparametric approximations for growth models with nonnormal random effects. *Multivariate Behavioral Research, 43*, 658–659.
Sterba, S., Prinstein, M. J., Bauer, D. J., & Cox, M. J. (2005, November). Unpacking heterotypic continuity: Gender and developmental factors differentiating latent trajectories of homotypic vs. heterotypic symptoms from 2–11 years. *Association for the Advancement of Cognitive and Behavioral Therapy*. New Orleans, LA.
Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York, NY: Wiley.
Tofighi, D., & Enders, C. K. (2007). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Greenwich, CT: Information Age.
van den Oord, E., Pickles, A., & Waldman, I. D. (2003). Normal variation and abnormality: An empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry, 44,* 180–192.
van der Vorst, H., Vermulst, A., Meeus, W., Dekovic, M., & Engels, R. (2009). Identification and prediction of drinking trajectories in early and mid-adolescence. *Journal of Clinical Child and Adolescent Psychology, 38,* 329–341.
van Dulmen, M., Goncy, E., Vest, A., & Flannery, D. (2009). Group-based trajectory modeling of externalizing behavior problems from childhood through adulthood: Exploring discrepancies in the empirical findings. In J. Savage (Ed.), *The development of persistent criminology.* Oxford Scholarship Online Monographs.
van Ryzin, M., Chatham, M., Kryzer, E., Kertes, D., & Gunnar, M. (2009). Identifying atypical cortisol patterns in young children: The benefits of group-based trajectory modeling. *Psychoneuroendocrinology, 34,* 50–61.
Verbeke, G., & Lessafre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91,* 217–221.
Verbeke, G., & Lessafre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis, 23,* 541–556.

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

Walton, K. E., Ormel, J., & Krueger, R. (2011). The dimensional nature of externalizing behaviors in adolescence: Evidence from a direct comparison of categorical, dimensional, and hybrid models. *Journal of Abnormal Child Psychology, 39,* 553–561.

Wiesner, M., & Kim, H. (2006). Co-occurring delinquency and depressive symptoms of adolescent boys and girls: A dual trajectory modeling approach. *Developmental Psychology, 42,* 1220–1235.

Xie, H., McHugo, G., He, X., & Drake, R. (2010). Using the group-based dual trajectory model to analyze two related longitudinal outcomes. *Journal of Drug Issues, 1,* 45–62.

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