DOI: http://dx.doi.org/10.53543/jeps.vol16iss4pp355-372

Latent Heterogeneity in High School Academic Growth: A Comparison of the Performance of Growth Mixture Model, Structural Equation Modeling Tree, and Forest

Amal Alhadabi
Sultan Qaboos University, Oman

Received: 2/10/2022 Accepted: 28/11/2022

Abstract: The Growth Mixture Model (GMM) is associated with several class enumeration issues. The contemporary advancement of automated algorithms presents two promising alternatives that merge confirmatory Structural Equation Modeling (SEM) with exploratory data-mining algorithms: SEM Tree and SEM Forest. This study investigated the performance of the aforementioned three methods (i.e., the GMM, SEM Tree, and SEM Forest) to detect latent heterogeneity in academic growth across four high school grades using an illustrative subsample of the Longitudinal Study of High School of 2009. The findings showed remarkable differences in detecting latent heterogeneity across the three methods as indicated by a parsimonious number of classes, with more unique growth trajectories, capturing the latent heterogeneity in the growth factors. In contrast, SEM Tree and SEM Forest were better at tracking the influences of covariates in the model parameters' heterogeneity, as indicated by providing more accurate measures of covariate importance and a detailed description of the role of covariates at each level of the tree or the forest. These findings imply the complementary use of these methods to obtain a clear separation between growth trajectories, as estimated by GMM; and the inclusion of most influential covariates, as identified by SEM Tree and Forest (208 words).

Keywords: academic growth, latent heterogeneity, growth mixture model, SEM tree, SEM forest.

التباين الضمني في النمو الأكاديمي لدى طلبة الثانوية: مقارنة بين نموذج الخليط المتعدد، وشجرة النمذجة البنائية، وغابة النمذجة البنائية

أمل الهدابي

جامعة السلطان قابوس، سلطنة عُمان

الملخص: يعد نموذج الخليط المتعدد من النماذج التقليدية التي تساهم في نمذجة التباين الضمني وفي ذات الوقت يعاني من إشكالية تحديد العدد الأمثل للمجموعات الضمنية الممثلة لهذا التباين. وتوفر التطورات المتسارعة في علم تنقيب البيانات الخوارزميات الأتوماتيكية التي تجمع بين النمذجة البنائية وتنقيب البيانات، وهما: شجرة النمذجة البيانية وغابة النمذجة البيانية. وتعدان طريقتان واعدتان في حل الإشكالية التي تواجه نموذج الخليط المتعدد. لذا هدفت الدراسة إلى المقارنة بين أداء ثلاث طرق إحصائية (نموذج الخليط المتعدد، وشجرة النمذجة البنائية، وغابة النمذجة البنائية) في تحديد التباين الضمني في النمو الأكاديمي لدى طلبة الثانوية باستخدام عينة توضيحية من الدراسة الطولية للمدارس الثانوية بالولايات المتحدة الأمريكية. أظهرت النتائج اختلافات جوهرية بين الطرق الثلاثة في نمذجة التباين الضمني حيث أظهر نموذج الخليط المتعدد أداءً أفضل من خلال تحديد عدد أقل من المجموعات الضمنية، ومسارات نمو فريدة التي توضح التباين في معامل النمو. في المقابل، شجرة النمذجة البنائية وغابة النمذجة البنائية أظهروا أداءً أفضل في تحديد أثر المتغيرات الديموغرافية في تباين معاملات النمو من خلال تحديد درجة أهمية كل متغير ودوره في كل تفرع للشجرة أو الغابة. وأشارت هذه النتائج إلى أهمية التكامل بين هذه النماذج للحصول على فرق واضح في مسارات النمو الذي يوفره نموذج الخليط المتعدد، وتضمين أكثر المتغيرات تأثيراً وأهمية الذي توفره شجرة وغابة النمذجة البنائية.

الكلمات المفتاحية: النمو الأكاديمي، التباين الضمني، نموذج الخليط المتعدد، شجرة النمذجة البنائية، غابة النمذجة البنائية

a.alhadabi2@squ.edu.om

Introduction

Modeling the latent heterogeneity in longitudinal data has been growing in popularity because it provides a more in-depth representation of the sample's underlying variability, forming several latent classes with unique growth trajectories and unbiased model parameters. The improved accuracy can be ascribed to modeling the extra variability between persons, which is considered a sizeable unexplained error variance when using other simple analyses (Preacher et al., 2008). Therefore, a broad range of disciplines have modeled latent heterogeneity in the growth of several developmental phenomena, such as educational outcomes (e.g., Alhadabi & Li, 2020) and psychological states (e.g., Infurna & Grimm, 2017), to name a few.

One of the more popular methods for capturing latent heterogeneity is the Growth Mixture Model (GMM). The GMM models the intra-person change over time (i.e., within-person growth), inter-person growth (i.e., between-persons variability), and the heterogeneity in the latent intercept and slope growth factors. In other words, the GMM achieves numerous tasks simultaneously, including (1) Estimating the initial status (i.e., intercept) and rate of change (i.e., slope), (2) Creating several latent classes with distinct growth factors, (3) Exploring the associations between the estimated classes and theory-driven selected covariates, and (4) Predicting distal outcomes by class membership (Muthén, 2008).

Despite the appealing implications of the GMM, mounting doubts have been recognized in the educational and psychological literature (e.g., Bauer, 2007; Nylund et al., 2007). These doubts can be attributed to the following: (1) Complexity of the class enumeration (i.e., identifying the correct number of classes), (2) Inconsistency among fit indices, and (3) Diversity of the factors influencing the GMM's accuracy. First, class enumeration is complex and follows a two-step approach. That is, several GMMs with an increasing number of classes are fitted (e.g., one-class, two-class, three-class, etc.). Then, a comparison of the fit indices is made to identify the best class structure (Ram & Grimm, 2009).

Second, several criteria are used to evaluate the optimal number of classes (e.g., Likelihood Ratio Tests [LRTs], Information Criteria [IC], entropy statistics, and multivariate skewness and kurtosis tests). Previous studies have produced contradictory conclusions under the same conditions (e.g., Grimm et al., 2014; Tofighi & Enders, 2007). For instance, Nylund and colleagues (2007) identified the Bayesian Information Criterion (BIC) as the best of the IC. In contrast, Li and Hser (2011) suggested that the BIC performs well only with large sample sizes and high separation.

Third, various factors influence class enumeration and latent heterogeneity deduction (e.g., class separation, sample size, mixing proportions, the inclusion of covariates, deviation from normality, etc.). Even though multiple prior studies have reported similar findings regarding the effect of some factors (e.g., sample size, class separation, and mixing proportions), a mixed bag of findings has been found for other factors (e.g., nonnormality and the inclusion of covariates). For instance, numerous studies have found that the inclusion of adequately chosen covariates increases class enumeration accuracy in the conditional GMM (e.g., Lubke & Muthén, 2007). On the other hand, other studies have found that the inclusion of covariates hindered class enumeration (Bauer, 2007; Tofighi & Enders, 2007).

The more recent data mining literature presents several automated algorithms that can capture the latent heterogeneity and alleviate some of the class enumeration issues associated with the GMM (McArdle & Ritschard, 2014). Two of these automated algorithms, which incorporate theory-guided structural equation modeling (i.e., SEM) and two data mining algorithms (i.e., decision tree and random forest), are SEM Tree (Brandmaier et al., 2013) and SEM Forest (Brandmaier et al., 2016). The SEM Tree algorithm models growth while recursively partitioning the dataset into nodes (i.e., classes) that maximally explain the differences in the model parameters, conditioning on observed covariates (Brandmaier et al., 2013).

Nonetheless, only a few studies have examined the performance of the SEM Tree in detecting latent heterogeneity (Jacobucci et al., 2017; Usami et al., 2017; Usami et al., 2019). Likewise, very little applied research has used the SEM Tree (Alhadabi, 2021; Brandmaier et al., 2013; Brandmaier et al., 2017). From these few studies, several technical issues have been discovered, including (1) divergence in the tree structure when using different control methods for selection bias and tree depth, (2) instability of the estimated classes, (3) dependency of the subsequent splits on the first split, and (4) computational burden (Brandmaier et al., 2016; Jacobucci et al., 2017; Strobl et al., 2009).

The SEM Forest, however, relieves some of the above-mentioned technical concerns associated with a single tree (Brandmaier et al., 2016). The advantage of using a forest (i.e., an ensemble of trees) is that it has the ability to stabilize the estimated classes and rank covariates based on their importance. Detailed investigation into how well the SEM Forest detects latent heterogeneity has received less attention in methodological studies. Also, a limited number of applied studies have utilized it (Brandmaier et al., 2017). To the best of the authors' knowledge, none of the applied studies to date have examined SEM Forest's performance in spotting heterogeneity when fitting the GMM.

Therefore, the current study aimed to compare the performance of three methods (i.e., the GMM, the SEM Tree, and the SEM Forest) using one of the illustrative a national data set. This empirical data set, the High School Longitudinal Study of 2009 (HSLS:09), addresses academic growth across four high school years (Ingels et al., 2011). The upcoming section provides an overview of the three methods and the measured heterogeneity in academic achievement.

Literature Review

Growth Mixture Model (GMM)

The GMM provides a parsimonious representation of between-person differences in within-person change over time, identifying various unobserved classes (Muthén & Muthén, 2000). Each latent class has unique values of latent growth intercept, slope, and variance-covariance parameters. Ignoring such unobserved heterogeneity leads to misleading information and incorrect growth parameters (Muthén & Shedden, 1999). Below is the mathematical notation for the measurement and structural parts of an unconditional GMM with c classes, as proposed by Muthén and Asparouhov (2008):

$$Y_{it}|C_{i=c} = \boldsymbol{\Lambda}_c \boldsymbol{\eta}_{ic} + \boldsymbol{\varepsilon}_i \qquad (1)$$

$$\boldsymbol{\eta}_{ij}|C_{i=c} = \boldsymbol{\alpha}_{ic} + \zeta_i \qquad (2)$$

Where Y_{it} is the $T \times 1$ vector of the measured outcome at T time points and $C_i = [1, 2, \ldots, k]$ is the $c \times 1$ vector of latent categorical classes. For a class c, Λ_c is the $T \times 2$ matrix of factor loadings in case of linear growth, and it extends to $T \times p$, where p is the number of latent growth factors when modeling nonlinear growth trends; η_{ic} is the $p \times 1$ vector of growth factors; ε_i is the $T \times 1$ vector of the error terms and is assumed to have a mean of zero and to be normally distributed. Furthermore, homogeneity of variance is assumed (i.e., homoscedastic across persons) at time T (Enders & Tofigi, 2008). Equation (2) describes each class's average value of the growth factors and the individual's variation around these average trajectories. Assuming a linear growth in class c, α_{ic} are the average intercept and linear slope. $\zeta_i = [\zeta_1, \zeta_2]$ is the individual's variation around the average class growth factors (i.e., variance-covariance matrix terms), which are assumed to be multivariate normally distributed with a mean of zero and a covariance matrix (Ψ). Time-invariant and time-varying covariates can be included. The measurement and structural parts of a conditional GMM are shown in Equations (3) and (4).

$$\boldsymbol{Y}_{it} \,|C_i = c = \boldsymbol{v}_c + \boldsymbol{\Lambda}_c\boldsymbol{\eta}_i + \boldsymbol{K}\boldsymbol{X}_i + \boldsymbol{\varepsilon}_i \qquad (3)$$

$$\boldsymbol{\eta}_{ij} \,|\mathrm{C}_i = \mathrm{c} = \boldsymbol{\alpha}_{jc} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}_{jc}\,\mathbf{X}_i + \zeta_{ij} \qquad (4)$$

As clarified in Equations (1) and (2), only additional terms are explained. $\boldsymbol{v}_c$ is the $T \times 1$ vector of measured initial status (i.e., observed intercept) at T time points for class c, which is constrained to zero for model identification (Bollen & Curran, 2006), $\boldsymbol{K}$ is the $q \times p$ matrix of regression coefficients, where q and p are the number of covariates and latent growth factors, respectively; and $\mathbf{X}_i$ is the $q \times 1$ vector of exogenous time-invariant covariates. In Equation (4), $\boldsymbol{\alpha}_{jc}$ is the $p \times 1$ vector of average latent factors in class c; $\mathbf{B}$ is the $p \times p$ matrix of correlation coefficients between the latent factors and the repeated measures; $\boldsymbol{\Gamma}_{jc}$ are the regression coefficients between covariates and growth factors; and $\boldsymbol{\zeta}_{ij}$ are the residual terms.

The class probability (i.e., the likelihood of classifying a person i in class c) is identified by a multinomial logistic regression model for k classes given a covariate X_i (Muthén & Asprouov, 2008), as shown in Equation (5).

$$P(C_i = c|X_i) = \frac{e^{(\alpha_c + \gamma_c' X_i)}}{\sum_{c=1}^{k} e^{(\alpha_c + \gamma_c' X_i)}} = \frac{e^{(\gamma_{0c} + \gamma_{1c} X_i)}}{\sum_{c=1}^{k} e^{(\gamma_{0c} + \gamma_{1c} X_i)}} \qquad (5)$$

$P(C_i = c \mid X_i)$ is the class probability. The γ_{1c} represents the increase in the log odds of being in class c relative to a normative group, given the covariate X_i. By default, the last latent class is considered the normative class (Muthén, 2008). The class probability is used to estimate the mixture distribution density in the GMM using Equations (6) and (7).

$$f(y_i|\Psi) = \sum_{c=1}^{k} \pi_c f_c(y_i|\theta_c) \tag{6}$$

$$\Psi = (\pi, \ \Theta) \tag{7}$$

Where π_c is the $c \times 1$ vector of the marginal distributions of k classes, which are also known as mixing proportions (McLachlan & Peel, 2000), where $0 \leq \pi_c \leq 1$, and the sum of all mixing proportions equal one. The f_c (y_i) is the $c \times 1$ vector of the conditional density function of y, given class membership that is assumed to be multivariate normal. The $\mathbf{\Psi}$ is a vector of unknown parameters in the model implied matrix (i.e., contain mixing proportions [$\pi_1, \ldots, \pi_{k-1}$] and the model parameters for each class [$\theta_1, \ldots, \theta_k$]; McLachlan & Peel, 2000).

SEM Tree

The SEM Tree is an automated model-based classification method that detects latent heterogeneity conditioning on significant covariates (Brandmaier et al., 2013). The classification outcome is a set of model parameters (e.g., intercept, slope, and variance-covariance matrix when fitting the growth curve model). The SEM Tree allows for the simultaneous achievement of three tasks: (1) modeling several SEM template models, (2) the selection of the most influential covariates that explain the variability in the model parameters, and (3) the classification of persons by partitioning the dataset into homogenous nodes (i.e., classes) in terms of model parameters (Brandmaier et al., 2014). In other words, the SEM Tree searches for the covariates that make the best split points, which results in a maximum within-class homogeneity (i.e., the most homogenous model parameters within a node) and greatest between-class heterogeneity (i.e., the most diverse model parameters among the nodes; Jacobucci et al., 2017). For instance, the SEM Tree may partition a dataset with two potential predictors (e.g., age and gender) based on one covariate or a combination of two covariates or neither depending on the strength of associations between the covariates and the model parameters (Brandmaier et al., 2013).

Following the mathematical notation adopted by Brandmaier and colleagues (2013), the SEM Tree algorithm has five steps. First, a pre-split template SEM model is estimated by a likelihood function $f(\theta|\mathbf{D}_F)$ for the whole dataset ($\mathbf{D}_F$: original learning dataset), where a vector of model parameters is calculated θ_F, as shown in Equation (8).

$$f(\theta|D_F) = -2\,\mathcal{LL}\,(\theta|\,\mathbf{D}_F) \tag{8}$$

Second, The full dataset is then split into non-overlapping subsets ($\mathbf{D}_1, \ldots, \mathbf{D}_k$).These independent subsets are used to test a compound model (i.e., the post-split model with nested nodes). The likelihood functions of the subsets are estimated independently. These functions are then summed to calculate the likelihood function of the compound model, as shown in Equation (9).

$$-2\mathcal{LL} = (T|\boldsymbol{D}) \sum_{\mathrm{d}\in\mathrm{D}} -2\mathcal{LL}(M(\theta_{\Psi(T,d)})|d) = \sum_{i=0}^{k} 2\mathcal{LL}(\hat{\theta}_i|\,D_i) \tag{9}$$

The $\mathbf{D}$ is a data set matrix and represented by $n \times (k + l)$, where n, k, and l are the sample size, repeated measurements, and un-modeled covariates, respectively. The l is not included in the observed dataset that is used to estimate the model parameters. That is, the SEM Tree creates two sub-datasets: (1) a dataset (i.e., $\boldsymbol{D}_k$) of repeated measures/observed variables that are used to estimate model parameters and (2) a covariate dataset ($\boldsymbol{D}_l$) that is used to match the person (d) to the node with the best fit sub-model, considering the person's covariate values using a mapping function [$\psi(T, \ d)$]. The mapping function traverses the tree with respect to the covariates. Meaning, it represents a process of visiting and checking each node across each covariate space, matching the person with the best sub-model to the correct node. When a nonsignificant covariate is found, the traversal process stops, terminating the classification process. Outweighing the issue of estimating local maxima in the GMM, the likelihood of observing the person within specific values of covariates in each node is maximal in the SEM Tree (Usami et al., 2017). There is no room for overlapping during the partitioning process because the estimation of model fit is conducted using one data set ($\boldsymbol{D}_k$). In contrast, the mapping function of the covariates is performed using another dataset ($\boldsymbol{D}_l$).

Third, a likelihood ratio test (Λ) is estimated, testing for the null hypothesis that the model parameters of the pre-split model equal the model parameters of the post-split model using Equation (10):

$$\Lambda = \sum_{i=0}^{k} 2\mathcal{LL}(\hat{\theta}_i |\, D_i) - 2\,\mathcal{LL}\,(\hat{\theta} |\, \mathbf{D}_F) \tag{10}$$

Values of Λ greater than a critical value (i.e., threshold) support the acceptance of the split, reflecting a maximum increase in the likelihood function after splitting, as inferred by a significant p-value. Fourth, steps 1 through 3 are repeated for all covariates and submodels. A process of numerous comparisons between the estimated probabilities is then executed, looking for the highest difference in the likelihood ratio to identify the most significant covariate at each splitting point.

This process may result in a long tree with many highly unstable, uninformative nodes which cannot be generalized (Hayes et al., 2015), considering the nature of the SEM Tree as a greedy recursive partitioning procedure (Brandmaier et al., 2013). Different methods have been suggested for controlling the tree depth and informativeness of estimated nodes. These control methods include: (1) prespecifying constraints or customized stopping criteria (pre-specified number of nodes, pre-specified number of participants per node, and prespecifying depth of the tree; Brandmaier et al., 2014; Usami et al., 2017), (2) applying the Bonferroni or cross-validation (cv) correction to control multiple comparisons, inflated Type I error, and selection bias (Usami et al., 2017), (3) applying one of the four maximum likelihood (ML) control methods that are known as pruning techniques (i.e., naïve, cv, fair, and fair3; Brandmaier et al., 2014; Jacobucci et al., 2017), and (4) a score-guided SEMTree.

The literature provides little guidance on which methods should be used. For instance, Jacobucci and colleagues (2017) state that prespecifying constraints are highly questionable. Furthermore, Brandmaier et al. (2013) suggested using ML control methods, precisely fair and cv. However, the precision of subsequent splits is accounted for by the first split's accuracy when using ML control methods (Grubinger et al., 2011). This means that failing to accurately estimate the first split results in a cumulative inaccuracy in the subsequent splits. Therefore, the score-guided SEM Tree may act as a remedy for multiple ML comparisons by proposing an additional five methods. One recent simulation study found that two score-guided methods (i.e., maxLMO and CvM) outperformed ML methods, as indicated by higher statistical power, reduced computational time, and better node recovery when examining multiple parameters (Arnold et al., 2020).

SEM Forest

The SEM Forest method is a hybrid of a theory-driven approach (i.e., SEM) and a data-driven approach (i.e., random forest). The random forest ranks and selects the influential covariates from a large set of covariates, considering the complex interactions among all covariates, resulting in a proximal clustering of cases adhering to model parameters (Brandmaier et al., 2016). The pre-specified SEM model is fitted to a trained dataset and validated on an out-of-bag dataset. The selection of the best model is based on the most informative covariates.

The mathematical notation presented by Brandmaier et al. (2016) states that for each tree in the SEM Forest algorithm with a size of t trees, a trained sample ($\mathbf{D}_i^{Train}$) and an out-of-bag sample ($\mathbf{D}_i^{OOB}$) are created. The $\mathbf{D}_i^{OOB}$ is used as a validation sample for the split decision estimated using the trained data, thereby strengthening the stability of estimated classes and ensuring the accuracy of the splitting decisions. This cross-validation approach provides realistic estimates of nodes that are expected to appear in new datasets (Hapfelmeier & Ulm, 2013). A subgroup of candidate covariates *(c)* is randomly chosen and analyzed by each tree to account for the covariates' variability. The literature has not produced a specific rule of thumb for the exact number of candidate covariates (i.e., *mtry*) to be tested. One option is to specify a fixed number of candidate covariates. Another option is to use an equation (e.g., $c = \sqrt{m}$, where m is the total number of covariates) to determine the number of candidate covariates (Breiman & Cutler, 2014).

Unlike the SEM Tree, the SEM Forest provides a visual presentation of two unique measures: Variable Importance (VI) and Case Proximity (CP). VI is a measure of model misfit when removing the covariate c from the sample (i.e., permutation importance). VI is estimated by following a four-step process (Brandmaier et al., 2016) as follows: (1) estimate a likelihood of model fit using an out-of-bag sample [$LL(\mathbf{D}_i^{OOB}|T_i)$], (2) create a new out-of-bag data by randomly removing covariate c values (i.e., scrambled $\tilde{\mathbf{D}}_i^{OOB}$), (3) estimate the second likelihood of model fit using the scrambled data and (4) calculate the differences in the misfit. The covariates are then ranked in descending order (i.e., from the highest to lowest misfit). Greater misfit (i.e., large differences in log-likelihood functions) indicates that the covariate is highly informative of the model parameters. Strobl et al. (2009) noted that small values around zero or even negative values illustrate low predictive values and low importance in identifying nodes. The CP quantifies the degree of closeness between each pair of cases in the sample (Breiman & Culter, 2014). Therefore, cases that are located in the same terminal node have a higher level of proximity. The proximity values are plotted with multidimensional scaling to measure the distance/dissimilarity between cases.

Academic Achievement in High School

Significant heterogeneity has been recognized in the academic achievement growth trajectories across the four high school years (Alhadabi & Li, 2020; Bowers & Sprott, 2012; Hodis et al., 2011; Lee & Rojewski, 2013; Muthén, 2008), resulting in the grouping of students into distinct latent achievement classes. This considerable heterogeneity is found in the starting values (i.e., intercept growth factor), the steepness of slope growth factor (i.e., the magnitude of change in GPA over time), and the shape of the change (i.e., linear or nonlinear).

The literature, however, does not agree on the number of latent classes. While some prior studies have found that two classes presented the unobserved variability well (e.g., Gottfried et al., 2017; Lee & Rojewski, 2013), other studies supported a three-class structure (e.g., Alhadabi & Li, 2020; Liu & Lu, 2011; Muthén, 2008;), and a four-class solution (Bowers & Sprott, 2012). As for growth shape, even though some prior research demonstrated a linear growth trajectory (Bowers & Sprott, 2012), other studies have supported a nonlinear trajectory (e.g., quadratic; Choi et al., 2016). Alhadabi and Li (2020) showed that the freely estimated nonlinear growth model had the best model fit.

A thorough review of the literature suggests that several individual-related and contextual-related covariates can explain this heterogeneity. The current discussion is limited to student-related covariates because this study investigated the performance of the simplest model, single-level GMM. At the student level, gender and socioeconomic status (SES) were the most influential covariates (Gottfried et al., 2017; Hodis et al., 2011; Lee & Rojewski, 2013; Muthén, 2008). Other covariates — including ethnicity, student locale, and negative student behavior — were also significant predictors but to a lesser extent (Alhadabi & Li, 2020; Bowers & Sprott, 2012; Muthén, 2008). For example, more female students and students with high SES were classified in the high-achieving class, while more male students and students with low SES were included in the low-achieving class (Hodis et al., 2011; Lee & Rojewski, 2013). Regarding ethnicity, the low-achieving class was more likely to have more African Americans (Alhadabi & Li, 2020, Muthén, 2008), while the mid-decreasing class had more Hispanics, and the high-achieving class had more Asian students (Bowers & Sprott, 2012).

Study Aim and Research Questions

As noted above, the literature has described various methods for modeling latent heterogeneity (i.e., the GMM, the SEM Tree, and the SEM Forest). Most of the published studies used the GMM (e.g., Gottfried et al., 2017; Hodis et al., 2011; Muthén, 2008). Jacobucci et al. (2017) examined the differences between finite mixture models and the SEM Tree. None of the prior studies compared the performance of the GMM and the SEM Forest. Furthermore, little is known about the performance of the SEM Tree and the SEM Forest in applied education research. This study, therefore, sought to compare the performance of three methods of modeling latent heterogeneity in academic growth; it aims to highlight the similarities and differences using illustrative data that assess academic achievement among a nationally representative sample of high school students in the United States.

The following questions were addressed:

1. Is there significant heterogeneity in the model parameters when modeling GPA growth? To what extent do the model parameters estimated by the three methods vary?

2. What is the best class structure when fitting GMM, and what is the number of estimated nodes when SEM Tree and SEM Forest are used?

3. What are the most influential covariates that explain the latent heterogeneity?

Methods

Illustrative Data

A random subsample of nationally representative high school students in the United States (N = 9,957) was taken from a public-use dataset, HSLS:09. The subsample was drawn from only those students who responded to the second follow-up phase of data collection (Duprey et al., 2018). This dataset was chosen because it assesses academic data for four successive years of high school, representing more recent data about Millenials' academic performance. The data set also includes several personal and contextual covariates that can affect students' academic performance (Ingels et al., 2011). The selected subsample contains complete data sets. A small subsample (i.e., n = 1,000) was plotted (see Figure 1) to visualize the growth in the dataset. The subsample showed diverse growth trajectories, with some scores increasing over time and others declining over time. The rate of change also was not constant.

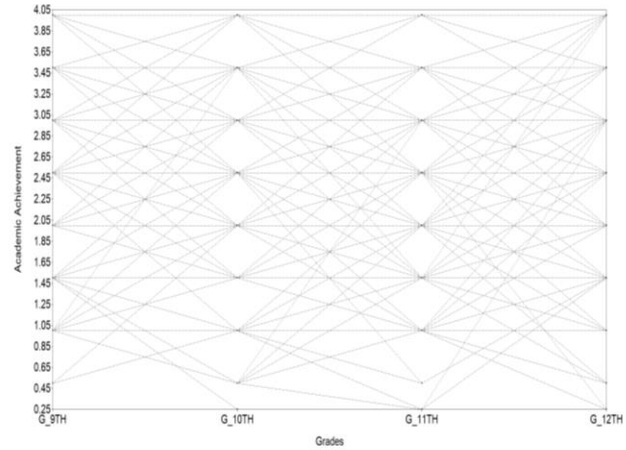

Figure 1. Trajectories of a random subset of 1000 students from 9th to 12th grades

Five covariates were examined: gender, ninth-grade SES, and three ethnic groups (i.e., White, Black/African American, and Hispanic). The sample had 49% males (n = 4,770) and 51% females (n = 5,187). Three ethnic groups were identified in the selected data: Whites (n = 7,536), Hispanic (n = 1,374), and African American (n = 1,451). The composite score of ninth-grade SES was computed using three variables: parents/guardians' education, family income, and the parents' occupations (Duprey et al., 2018). The mean SES of the sample was .12 (SD = .79).

Data Analysis

Descriptive statistics were assessed using R (R Core Team, 2020). A preliminary inspection of the data was conducted (e.g., missing data, normality, and outliers). This section reviews the specification of three methods (i.e., the GMM, SEM Tree, and SEM Forest; see codes in supplementary material).

Specification of the GMM

A two-step approach was adopted to fit the GMM, modeling the growth trajectories using Mplus 8 (Muthén & Muthén, 2017). In Step 1, multiple unconditional class-invariant variance and covariance models (GMMs-CIs) with nonlinear growth and an increasing number of classes were fitted to identify the best class-structure (Ram & Grimm, 2009). Justification for fixing the variance-covariance matrix across latent classes came from Alhadabi and Li (2020). They analyzed a large sample of HSLS:09 (N = 12,314) and found that GMM-CI with a freely estimated slope (i.e., capturing the nonlinearity) had the best model fit. In Step 2, a conditional model was fitted, identifying the most influential covariates (see Figure 2). The current study modified the multiple default options, which were: (1) STARTS = 400, 100, (2) LRTSTARTS = 0 0 100 25, and (3) LRTBOOTSTRAP = 50. These modifications were conducted to fix local maxima and improper solutions, consistent with Hu et al. (2017).

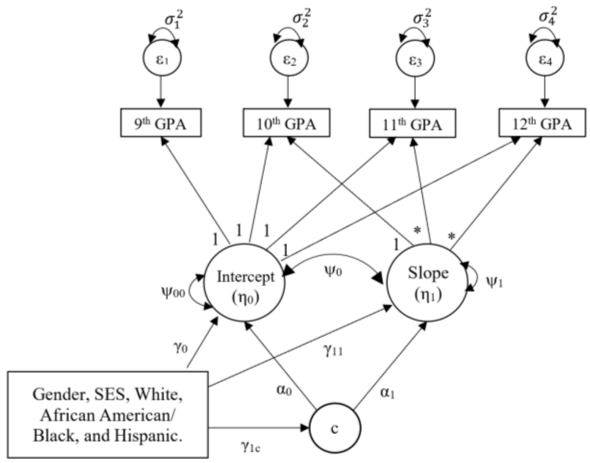

Figure 2. The conditional GMM

Several indices were evaluated, which include: (1) AIC, BIC, and SABIC from the IC, (2) LMR, VLMR, and BLRT from the LRTs, and (3) Entropy. Overall, a model with the lowest IC heralds a good fit (Tofighi & Enders, 2007). Related to LRTs, a significant *p*-value suggests acceptance of the k-profile model and a rejection of the *k*-1 model, where *k* is the number of latent classes. Entropy values of 0.80, 0.60, and 0.40 indicate high, medium, and poor classification, respectively (Muthén, 2008). Other criteria were considered, including (1) average posterior probabilities (i.e., should be near one and ≥ 0.70; Wang & Wang, 2012), (2) profile size (i.e., < 5% was dismissed), (3) parsimony of the model, and (4) interpretability of estimated class-structure (Berlin et al., 2014).

Specification of the SEM Tree

The SEM Tree analyzed a GMM following four steps: (1) Create the SEM model using the OpenMx package in R (Boker et al., 2011; Grimm et al., 2017), (2) Run a single tree with no control method using the semtree package (Brandmaier, 2015), (3) Create a control object, and (4) Rerun the tree. In the first step, a latent growth model (i.e., assuming a single class) was run. The reason for fitting the latent growth model instead of GMM was an error message that stops running the tree when fitting GMM directly. Jacobucci et al. (2017) used a similar approach when fitting finite mixture models, thus lending support to the decision made in this study.

A global constraint was specified in the second step, which states that four error variance terms and within-class variance-covariance terms were set to be equal across the nodes, aligning with the GMM specification. Then, the tree was run without applying any control methods, which maximally explain the differences in the model parameters (i.e., intercept and slope growth factors) conditioning on influential observed covariates. In the third step, three options were specified to control the tree depth, including (1) the max depth of the tree (i.e., max.depth = 3), (2) specifying score-guided control method to avoid the pitfalls of ML control method, and (3) controlling for multiple comparisons using the Bonferroni method. These options aligned with some prior studies (Arnold et al., 2020; Brandmaier et al., 2013; Jacobucci et al., 2017). A visual presentation of the tree and the growth trajectories for each node were obtained.

Specification of SEM Forest

Five steps were followed to specify the SEM Forest, including (1) creating the model, (2) running a single tree with no control method, (3) creating a control object and rerun the tree, (4) creating a control object for the SEM Forest, and (5) creating a forest to visualize the covariates' importance and estimate case proximity (Brandmaier, 2015). The specification of the first three steps was identical to those used in for the SEM Tree as stated above. In the fourth step, four options were specified to create a control object, including (1) num.trees = 30 (i.e., representing the total number of trees in the forest), (2) SEM forest.control = fair (i.e., controlling for selection bias of covariates and splitting points), and (3) mtry = 2 (i.e., defining the number of covariates that are randomly selected at each splitting point to estimate variable importance). In the fifth step, SEM Forest was run. Next, the variable importance measure was plotted. In addition, a case proximity matrix (P) was created, and principal component analysis (PCA) was conducted to project the cases across a two-dimensional plot and facilitate the interpretation of clusters in the proximity matrix.

Results

Descriptive statistics and correlations. Table 1 summarizes the descriptive statistics indicating that normality was fulfilled. The Zero-order Pearson correlation coefficients between the variables were estimated. These coefficients reflected significant associations between the studied variables.

GMM Results

The findings of several unconditional GMM-CI models showed that the three-class, four-class, and five-class models were candidates to capture the latent heterogeneity accurately (see Table 2). The four-class and five-class models had the lowest AIC, BIC, and SABIC. They also had higher entropy and significant LMR-LRT and BLRT. However, the class sizes and average posterior probabilities were smaller than the optimal range (i.e., < 5% for Class 2 in the four-class and five-class GMM), suggesting a deficient classification. In addition, several convergence issues appeared despite altering the default setting, as summarized in the methodology. In contrast, the three-class model had similar smaller values of AIC, BIC, and SABIC. The class average posterior probabilities and class sizes were optimal. These results suggest that the three-class GMM-CI had the best model fit and had the most accurate class enumeration.

On average, students in Class 1 were low achievers in ninth grade (η_I = 2.08, p < .001, n = 1,727) and showed non-significant GPA decline over time ((η_S = -.05, p = .06). This class was named "Low-achievers" (see Figure 3). Students in Class 2 started with moderate academic achievement in ninth grade (η_I = 2.29, p < .001, n = 950) but increased their GPA during the four years of high school (η_S = .05, p < .01). Thus, Class 2 was named "Moderate growing-achievers and". Lastly, students in Class 3 were relatively highly achievers with an average intercept (η_I = 3.29, p < .001, n = 7,280) and their performance remained constant over time (η_S = .003, p = .06). This class was named "High-achievers". Significant variability was identified in the intercept (ψ_{00} = .22, p < .001). However, the slope variance (ψ_{11} = .00, p = .21) and covariance were not significant (ψ_{01} = .003, p = .06).

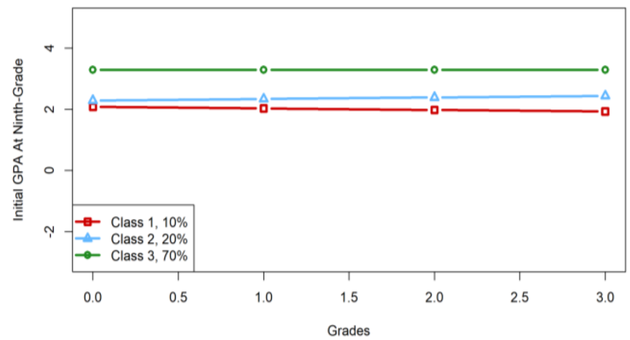

Figure 3. Three-class GMM-CI model.

Table 1. Pearson Correlation Coefficients between the Selected Variables (N = 9,957)

Variables	1. 9th-grade GPA	2. 10th-grade GPA	3. 11th-grade GPA	4. 12th-grade GPA	5. Gender	6. Hispanic	7. White	8. Black	9. 9th-grade SES
M	2.97	2.94	2.95	3.05	1.52	.14	.76	.15	.17
SD	.78	.78	.76	.74	.50	.34	.43	.35	.79
S	-.60	-.60	-.65	-.84	-.08	2.10	-1.20	2.01	.29
K	-.20	-.16	.06	.60	-1.99	2.41	-.57	2.03	-.21
1	-								
2	.81***	-							
3	.72***	.79***	-						
4	.65***	.69***	.74***	-					
5	.15***	.16***	.15***	.18***	-				
6	-.15***	-.14***	-.12***	-.11***	.01	-			
7	.07***	.09***	.07***	.10***	-.01	-.01	-		
8	-.21***	-.21***	-.18***	-.19***	.02	-.02**	-.52***	-	
9	.35***	-.34***	.33***	.30***	-.02**	-.24***	.07***	-.13	-

Note. *p < .05, **p < .01, ***p < .001.

Table 2. GMM_CI Models with Freely Estimated Slopes

Fit Statistics	Two-class	Three-class	Four-class	Five-class
A. GMM-CI				
LL(No. of parameters)	-30633.25(14)	-30314.25(17)	-30129.71(20) a b	-30041.76(23)a b
AIC	61294.50	60662.49	60299.41	60129.53
BIC	61395.39	60784.99	60443.53	60295.26
SABIC	61350.90	60730.97	60379.97	60222.17
Entropy	.82	.72	.75	.74
LMR-LRT(p)	1059.52(.00)	615.72(.00)	356.18(.00)	169.74(.00)
BLRT(p)	-33182.20(.00)	-30633.25(.00)	-30314.24(.00)	-30129.71(.00)
Group size (%) C1	9027(91%)	950(10%)	956(10%)	581(6%)
C2	930(9%)	1727(20%)	435(4%)	382(4%)
C3		7280(70%)	7166(72%)	5365(54%)
C4			1400(14%)	2987(30%)
C5				642(6%)

Note. [a] small class's probabilities, [b] convergence issues (e.g., negative variance, the likelihood has not replicated and local maxima)

The finding of a conditional three-class GMM_CI (see Table 3) showed that Hispanic, White, and African American were negatively associated with the intercept growth factor in three classes. In contrast, gender and ninth-grade SES positively correlated with latent intercept, implying females and students with a higher SES had higher 9th-grade GPA. In comparison, only three covariates (i.e., White, African American, and SES) had significant associations with the latent slope in three classes. White and African Americans negatively correlated with latent slope, indicating that the GPA of White and African Americans declined over time. However, students with higher SES showed positive GPA growth over time. The other two covariates (i.e., gender and Hispanic) did not significantly correlate with GPA change over time, suggesting that students in three classes had a similar GPA growth regardless of gender and ethnicity.

The multinomial logistic regression coefficients reflected the likelihood of belonging to a specific latent class relative to a normative class, the "High-Achievers" class. Odds were estimated for the significant coefficients (see Table 3.C). The odds of classification in "Low-Achievers" class increased by (1) .70 when students were male, (2) 1.53 when students were African American, and (3) .45 when the students had a low SES. Most students in the "Moderate-Growing Achievers" class were male, African American, and students with low SES.

SEM Tree Results

The SEM Tree found that 9th-grade SES, Black/African American, and gender were the most influential covariates in classifying students based on model parameters' heterogeneity (see Figure 4.a). The tree resulted in six splitting points and seven nodes. Overall, all nodes had different starting values, and they had ascended to constant growth over time. In other words, no major differences were obtained in the magnitude and the direction of the slopes (see Figure 4.b).

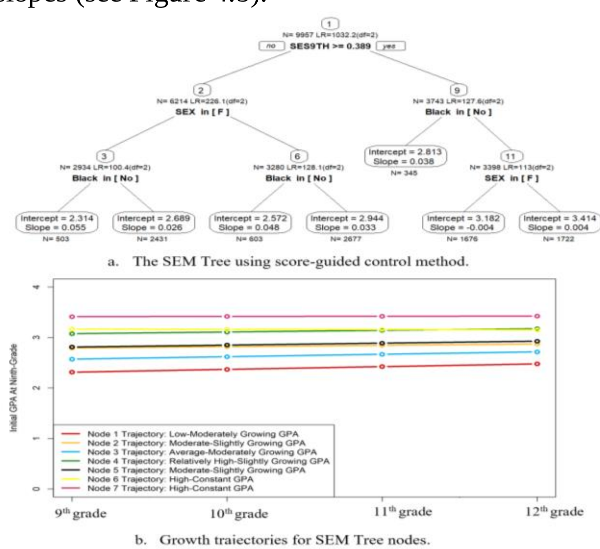

Figure 4. Modeling growth using SEM Tree

Note. Black in [NO] = If No, it means Black; if Yes, it means non-Black. SEX in [F]= If No, it means male; if Yes, it means female.

Table 3. Multivariate Standardized Regression and Multinomial Logistic Regression Estimates of Covariates on the Latent Growth Factors and Class Membership

	Low Achievers Class		Moderate-growing Achievers Class	High Achievers Class	
	Intercept latent factor				
Gender	.19***		.19***	.19***	
Hispanic	-.11***		-.10***	-.09***	
White	-.08***		-.07***	-.07***	
African American	-.27***		-.28***	-.22***	
9th-grade SES	.22***		.23***	.26***	
	Slope latent factor				
Gender	-.03		-.03	-.03	
Hispanic	-.02		-.02	-.02	
White	-.13***		-.12***	-.12***	
African American	-.11***		-.12***	-.09***	
9th-grade SES	.05***		.05***	.06***	
	Class Membership where High Achievers is a Reference Class.				
	Low Achievers Class			Moderate-growing Class	
	Coefficient		Odds	Coefficient	Odds
Gender	-.36***		.70	-.59***	.56
Hispanic	-.17			-.16	
White	-.03			.18	
African American	.43*		1.53	.66***	1.95
9th-grade SES	-.80***		.45	-.86***	.42

*Note***.** *p < .05, **p < .01, ***p < .001.

In detail, the 9th-grade SES was the most influential covariate, resulting in the first split that classified students into low SES (i.e., < .39) and high SES (i.e., ≥ .39) groups. Among low SES students, gender and Black/African American created the second and third split, respectively. These splits resulted in the formation of four nodes/latent classes: (1) Low-Moderately growing GPA, (2) Moderate-slightly growing GPA, (3) Average-Moderately growing GPA, and (4) Relatively high-slightly growing GPA.

Students in the first class (i.e., Low-Moderately growing GPA; n = 506) were males, Black, and had low 9th-grade SES. These students had the most inferior initial status and showed a moderate growth as indicated by the highest slope compared with other nodes (i.e., η_s = .06). Students in the second node, “Moderate-Slightly Growing GPA”, were males, non-Black with low SES (n = 2,431). They started with moderate initial GPA status at 9th grade and demonstrated a slight growth. The third node’s students (i.e., Average-Moderately growing GPA; n = 603) were Black, females, and had low SES. They started with an average 9th -grade GPA and showed comparable growth with their male peers in the first node. In the fourth node (i.e., Relatively High-Slightly Growing GPA; n = 2,677), students had a near 3.0 GPA and demonstrated slight growth over time. These students were female, non-black, and with low 9th-grade SES.

In contrast, Black and gender made the second and third split among students with high 9th-grade SES (i.e., > .39), respectively, which resulted in classifying these students into three nodes: (1) Moderate-slightly growing GPA, (2) High-constant GPA, and (3) High-constant GPA. Students in “Moderate-slightly growing GPA” node were Black students with high SES (n = 345). They had GPAs below 3.0 and demonstrated a slight growth over time (i.e., η_s = .04). The other two nodes started with a high 9th-grade GPA (i.e., above 3.0) and maintained this steady performance across time. One of them consisted of non-Black males with high SES, while the students in another node were non-Black females with high SES.

SEM Forest Findings

The findings of the first SEM Forest measure (VI) showed that the 9th-grade SES was the most influential covariate, as indicated by an average absolute increase of -2LL of 1809.55, reflecting the increase in the misfit when SES was permuted from the data (see Figure 5). Gender and Black were the second and third informative covariates with a relative increase in the misfit of 362.34 and 333.26, respectively. These findings substantiate the nodes' structure estimated by SEM Tree. As suggested by Brandmaier and colleagues (2016), a single tree was run to identify the best split point in the continuous covariate (9^{th}-grade SES). The finding showed that a split point at .36 was the most significant level of 9^{th}-grade SES. Then, the SES was recoded to a dichotomous variable (i.e., low and high SES) to facilitate the interpretation of the case proximity measure.

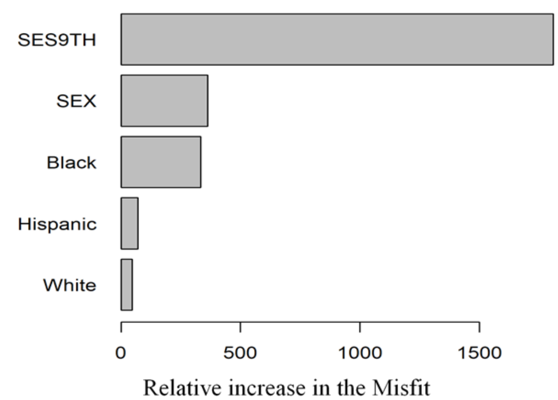

Figure 5. SEM Forest's VI measure.

The second measure, a case proximity matrix (***P***), was created for each pair of the sample (N = 9,975). A dissimilarity matrix was created (***P*** -1; Brandmaier et al., 2016). Then, a principal component analysis (PCA) was conducted using the obtained matrix to project the cases across a two-dimensional plot (see Figure 6). The findings showed a two-cluster structure. The principal vertical axis accounted for 39.02% of the variance, expressing a range of two levels of 9^{th}-grade SES: Low SES (i.e., presented by black color) and high SES (i.e., showed by red color). The principal horizontal axis explained 35.52% of the data's variance based on gender. The dashed vertical line split the data based on gender, resulting in a two-cluster structure: Male (i.e., left side) and female (i.e., right side). The shape illustrates the differences based on the being non-black (i.e., clusters of circles in the left and right side across the horizontal line) and Black (i.e., triangles in the middle area across the horizontal line).

The case proximity showed the presence of eight sub-clusters that had unique participants. A latent growth model was fitted for each cluster (see Table 4). The findings showed that the clusters had unique growth trajectories (see Figure 7). Overall, several observations were noted. Groups with high 9th-grade SES had higher initial GPA and demonstrated better growth compared to students with low SES. When holding SES and gender constant (e.g., male who had low SES), the clusters with Black students started with lower initial status (ηI = 2.32) com-pared to non-Black students (ηI = 2.69); however, they had relatively higher growth across time. When fixing the SES and ethnicity (e.g., High SES and Black), the clusters that contained female stu-dents had higher initial status and better growth over time than males.

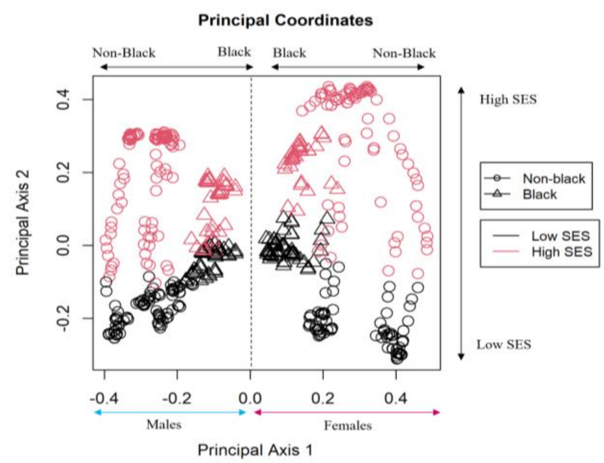

Figure 6. Case proximity measure.

Note. The color theme represents two levels of the 9^{th}-grade SES: low (black) and high SES (Red). Ethnicity was presented by different shapes: Black (triangle) and non-Black (circles). The dashed line split the data based on gender: Male (left side) and females (right side).

Like the nodes obtained from SEM Tree, these clusters were: (1) Cluster 1 contains Blacks males with low SES and was named as a low-moderate growing cluster (n = 499), (2) Cluster 2 had males, non-Blacks, had low SES (n = 2,384) and named as "Average-slightly growing GPA", (3) Cluster 3 started with average GPA at 9th grade and showed moderate growth across time, containing Black females with low SES (n = 598), (4) Cluster 4 had a near 3.0 GPA at 9th-grade and had an average increase across time, comprising non-Black females with low SES (n = 2,638), (5) Average-moderately growing cluster that contains Blacks males with high SES (n = 164), (6) High-slightly decreasing cluster (n = 1,723) where students were non-Blacks males with high SES, (7) High-slightly growing cluster that comprises Blacks females with high SES (n = 190), and (8) High-constant growing cluster that contains non-Blacks females with high SES (n = 1,761).

Table 4. Growth Parameters for the Subgroups Derived from the SEM Forest

Description	Low SES				High SES			
	Male		Female		Male		Female	
	Black (*n* = 499)	Non-Black (*n* = 2,384)	Black (*n* = 598)	Non-Black (*n* = 2,638)	Black (*n* = 164)	Non-Black (*n* = 1,723)	Black (*n* = 190)	Non-Black (*n* = 1,761)
Model Parameters								
Intercept	2.32	2.69	2.57	2.94	2.63	3.17	2.98	3.41
Slope	.05	.03	.05	.04	.04	-.004	.03	.01

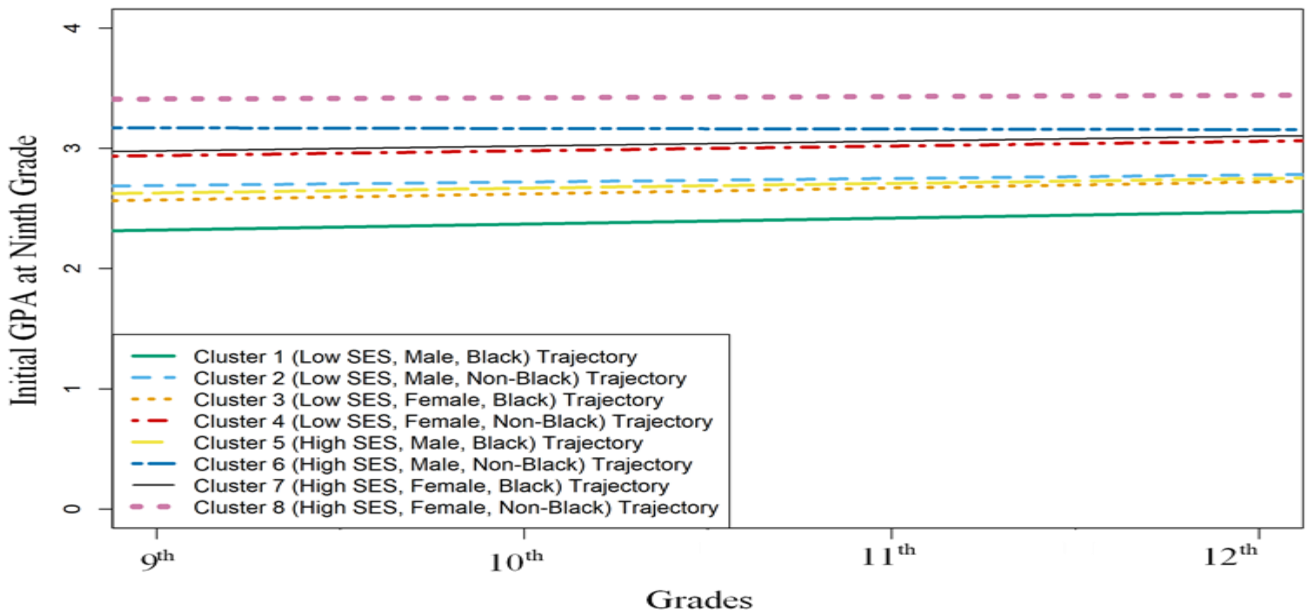

Figure 7. Growth trajectories for SEM Forest clusters.

Methods Comparison Findings

Several difficulties were faced to obtain a common criterion to compare between the confirmatory GMM and the data mining methods because they adopt different approaches in classifying persons. That is, GMM tends to form the classes based on estimating maximum likelihood (i.e., probabilistic approach) that maximizes the heterogeneity between classes in terms of growth factors whether covariates are included or excluded. In contrast, SEM Tree and SEM Forest follow greedy, top-down, recursive partitioning procedures that stop splitting the data when no important covariate is identified (i.e., all-or-nothing; Brandmaier et al., 2013). Five comparison criteria were examined: (1) The number of estimated classes, (2) The heterogeneity in the model parameters, (3) the covariates importance, and (5) Strengths and drawbacks, aligning with Jacobucci and colleagues (2017).

In detail, the number of estimated latent classes varied across GMM (i.e., three classes), SEM Tree (i.e., seven classes), and SEM Forest (i.e., eight classes). Related to heterogeneity in the model parameters, clear separation in the latent intercept factor was observed across three classes estimated by GMM. In contrast, less breakup and greater homogeneity in the intercept were noted across the classes obtained by SEM Tree and SEM Forest. GMM produced latent classes with more distinct patterns of change across time (i.e., latent slope factor), whereas the classes estimated by SEM Tree had similar growth trends. In between, SEM Forest produced more unique growth trajectories compared to SEM Tree. These observations can be attributed to the approach followed by these methods in the formation of classes. Meaning, GMM tends to form classes based solely on the heterogeneity of growth factors, resulting in more clear differences in the growth factors across the classes. However, data mining algorithms (i.e., SEM Tree and SEM Forest) create classes based on the growth factor heterogeneity conditioning on influential covariates (Brandmaier et al., 2013; Brandmaier et al., 2016). That is, the data split accounts more for covariates' importance.

Around covariates importance, the results from three methods confirmed that 9th-grade SES was the most significant covariate. Related to the second and third levels, while GMM found that Black was the second most influential covariate, SEM Tree revealed gender and Black were interchangeably important as second influential covariates. That is, SEM Tree provided a more detailed description of the influence of these covariates. For example, gender was at second rank in terms of importance among students with low SES, followed by ethnicity as Black/African American. In contrast, Black was the second important covariate among students with high SES. SEM Forest aligned with SEM Tree by giving more weight to gender as a second most crucial covariate, resulting in a more remarkable misfit when permuted. This finding contradicts GMM findings that identified gender as the third important covariate based on the magnitude of the regression coefficient. Outweighing the detailed description presented by SEM Tree, GMM provided two sets of regression coefficients, gauging the influence of covariates on the intercept and slope growth factors separately.

The three methods had several strengths and drawbacks, implying the complementary use of these methods. That is, GMM requires manual setting of the expected number of classes, estimating of multiple fit criteria, and conducting a comparison between models with an increasing number of classes when fitting unconditional model. Further specification is needed after identifying the best class enumeration by examining the effect of covariates when fitting conditional GMM. Data mining algorithms, in contrast, require no prior specification of the number of classes and no input related to the influence of covariates. In other words, SEM Tree and Forest automatically identified the associations between model parameters and covariates.

However, SEM Tree and Forest had many drawbacks. First, not all SEM models can be directly fitted using these methods. For example, this study found that GMM cannot be fitted directly, as indicated by obtaining an error message that prevents running the tree. Meaning, the tree failed to run the model when analyzing a mixture of classes. Second, the findings of these algorithms depended on the selected control method. That is, various control methods resulted in different tree-structure, the magnitude of covariates splitting points, and the number of nodes. Little methodological guidance was found related to the specification of these methods, suggesting conducting more simulation studies investigating the influence of the control method would be highly valuable. Third, the computation burden of the forest was enormously large, mainly when the sample size was large (i.e., $N > 5{,}000$). The resulted proximity matrix was huge in which the device failed to analyze. Correspondingly, a smaller number of trees was used to alleviate this issue.

Discussion

Modeling latent heterogeneity provides a more comprehensive examination of latent growth factor variability when analyzing longitudinal data. As a result, several latent classes are created with accurate model parameters reflecting distinctive growth trajectories. The GMM, a traditional probabilistic approach of modeling this heterogeneity, had several well-known concerns. Recent development in the literature presents two new analytic alternatives (i.e., SEM Tree and SEM Forest) that merge structural equation modeling and data mining algorithms. Therefore, the current study compared the performance of these three methods (GMM, SEM Tree, and Forest) in modeling the latent heterogeneity using one of the illustrative national data set (HSLS:09).

The unconditional GMM found that the three-class structure had the best fit, aligning with previous research findings (Alhadabi & Li, 2020; Muthén, 2008). The conditional GMM found that five covariates (i.e., gender, Hispanic, White, African American, and ninth-grade SES) had significant effects on the intercept growth factor. In contrast, only three factors (i.e., White, African American, and ninth-grade SES) were significantly associated with slope growth factor. These findings align with Alhadabi and Li (2020) findings, who found more and stronger influences of studied covariates on the intercept growth factor in their research. Simultaneously, they revealed limited and weaker effects of covariates on slope growth factor.

In detail, Hispanic, White, and African Americans had lower 9th-grade GPA compared to other students in three classes, while females and students with a higher SES had higher 9th-grade GPA, consistent with previous studies (Alhadabi & Li, 2020; Gottfried et al., 2017; Hodis et al., 2011; Muthén, 2008). Related to slope growth factor, White and African Americans showed negative growth (i.e., declining slope), whereas students with high SES had a significant positive increasing rate of change compared to other students. These findings substantiate Bowers and Sprott (2012) findings, which revealed that students with high SES showed greater academic growth over time among the high-achievers class. Most students in the "Low Achievers" class were male, African American, and had a low ninth-grade SES. Similarly, the odds of classification in the "Moderate-Growing Achievers" class significantly increased when students were male, African American, and had a low ninth-grade SES. In contrast, most students in the "High Achievers" class were female, non-African American, and students with high ninth-grade SES.

The SEM Tree findings revealed that a seven-class structure was the most reasonable structure in capturing the heterogeneity in growth factors conditioning on the influential covariates, contradicting previous studies, which estimated the class structure using GMM (Liu & Lu, 2011; Muthén, 2008). Like GMM, these significant covariates were 9th-grade SES with the best split-point of .39, gender, and Black/African American. The estimated classes had similar values of initial GPA status and demonstrated similar rising to constant growth trajectories. Unlike GMM, less clear heterogeneity in the growth factors was observed between classes. Because the formation of classes was not made solely on growth factors, but rather it was made by conditioning on influential covariates, this adds support to the findings of prior studies (Jacobucci et al., 2017). Like GMM, the class that contains Black males with low SES had the lowest initial 9th-grade GPA and showed the highest growth over time. When fixing the SES level, females had a higher 9th-grade GPA compared to their male peers. Overall, non-Black students had higher initial status than Black students. Nevertheless, Black students demonstrated a positive and greater rate of change across time compared to non-Black. The estimated influences of examined covariates validate the findings of prior studies (Hodis et al., 2011; Lee & Rojewski, 2013).

SEM Forest provided greater support to an eight-class structure compared with GMM and SEM Tree. In a continuum between GMM and SEM Tree, SEM Forest showed intermediate heterogeneity between classes in the growth factors. Like GMM and SEM Tree, SEM Forest found that SES was the most influential covariate with a slightly different splitting point (i.e., SES < .36 and SES ≥ .36), where high SES classes had higher 9th-grade GPA and showed relatively smaller to constant slope compared with low SES classes. Unlike SEM Tree, SEM Forest placed gender second in importance, followed by Black. The classes containing females had better average initial performance compared to males when fixing SES and ethnicity. When holding SES and gender constant, the clusters with Black students started with lower initial status than non-Black students; though, they had a relatively higher rate of change across high school years, extending the findings of Alhadabi and Li (2020).

Implications and Limitations

This study resulted in constructive conceptual and methodological findings. The current study demonstrated significant variability in academic growth among high school students, suggesting that treating students as a homogenous group may result in biased generalization. Sources of this variability imply the inherent effect of circumstantial variables (SES, ethnicity, and gender) beyond students' academic abilities. For instance, despite Black/African American males with low SES having the lowest average initial status, they showed relatively greater growth across high school. This study also highlighted the characteristics of students that were classified in each class, providing data-driven recommendations for conducting targeted interventions to empower students' academic performance. These interventions and initiatives are not necessarily limited to the student level; they can be oriented to school administrators and educators. For instance, targeted initiatives are needed for students in low-achiever classes, particularly in communities and school districts where the majority of students have low SES in Black/African American communities.

This study articulated many essential methodological differences among the three methods in detecting latent heterogeneity. The GMM resulted in a sparse number of classes with good separation in growth factors, whereas the SEM Tree and Forest produced plentiful classes with nearly similar growth parameters. Second, the SEM Tree and Forest were superior to GMM in providing a detailed description of the effects of the covariates at each level of the tree. These findings emphasize the importance of the complementary implementation of the three methods. Meaning the SEM Tree and Forest can be used in a preliminary step to identify the most influential covariates, mainly when the data contain a large number of explanatory variables. Subsequently, an unconditional and conditional GMM with the most influential covariates can be used to model the latent heterogeneity.

In addition, GMM is a confirmatory approach, while SEM Tree and Forest are exploratory data-driven algorithms. The dilemma between these confirmatory and exploratory analyses is crystal clear in the literature. There is an obvious preference for developing well-articulated hypotheses that should be modeled using confirmatory techniques, which hypothetically leads to substantial findings with small margins of error (Schumacker & Lomax, 2016). In practice, researchers conduct simple and complex statistical analyses in a more exploratory style, specifically in the case of statistical non-significance or poor model fit (McArdle & Ritschard, 2014). With the plethora of available big data and statistical development, the importance of mutually implementing these confirmatory and exploratory analyses to answer more profound and diverse research questions is becoming undeniable. This study aimed to bridge the divide between these methods by highlighting the similarities and differences in detecting latent heterogeneity.

This study also has multiple limitations. It investigated a limited number of student-related covariates, which does not reflect the abundance of explanatory covariates at the family and school levels because the public version of HSLS: 09 suppresses many contextual variables and identifiers of schools, resulting in the exclusion of the higher level of nested data. The current study used a score-guided control method to adjust the tree depth. Nevertheless, the tree structure may vary considerably based on the selected control method (Hayes et al., 2015). This research also used a small number of trees (i.e., 30 trees) in the forest to manage the computational burden associated with a large sample size.

Based on the findings, the current study has several recommendations. First, educational researchers should use a multilevel growth mixture model to examine the effect of contextual covariates on academic heterogeneity and account for the hierarchical structure of the national data. Second, methodological researchers should develop a new automated unsupervised algorithm that merges the strengths of GMM and SEM Forest. Further examination of the new algorithm with these three methods is suggested by conducting a simulation study that examines different experimental conditions. Lastly, based on the present study, the authors strongly advocate for performing a simulation study investigating the effects of various control methods in the performance of the SEM Tree and Forest under several design factors.

Conclusion

To confirm the differences between the sets of early and late-acquired words, the researchers conducted a paired In conclusion, detecting the latent heterogeneity in longitudinal educational and psychological attributes leads to accurate and unbiased estimates, which enrich the literature with precise findings and valuable recommendations. Several methods can be used to detect this heterogeneity. The current study investigated the performance of the traditional estimation method (i.e., ML/EM) and the two data-mining automated estimation methods (i.e., SEM Tree and Forest). This study attempted to build bridges between the confirmatory and exploratory approaches by highlighting the similarities and differences.

The findings found remarkable differences in the performance of the three methods. The scale of positive performance was tilted toward GMM, as indicated by forming classes with more unique growth trajectories, and more accurately capturing the latent heterogeneity in the growth factors. SEM Tree and Forest, in contrast, performed better in tracking the influences of covariates on the model parameters' heterogeneity, as indicated by providing more accurate measures of the covariates importance. The findings emphasized that the performance of the three methods was not totally optimal. In the end, the current study recommends the complementary implementation of these methods to obtain a clear separation between growth trajectories, as estimated by GMM, and the inclusion of the most influential covariates, as identified by the SEM Tree and Forest.

Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alhadabi, A. (2021, April). *Science Interest, Utility, Self-Efficacy, Identity and Science Achievement among High School Students: An Application of SEM Tree.* Paper presented at the meeting of the American Educational Research Association, Orlando, USA.

Alhadabi, A., & Li, J. (2020). Trajectories of academic achievement in high schools: Growth Mixture Model. *Journal of Educational Issues, 6*(1), 140-165.

Arnold, M., Voelkle, M., & Brandmaier, A. (2021). Score-Guided Structural Equation Model Trees. *Frontiers in psychology, 11*(564403), 1-18.

Bauer, D. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research, 42(*4), 757–786.

Berlin, K., Parra, G., & Williams, N. (2014). An introduction to latent variable mixture modeling (Part 2): Longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology, 39*(2), 188–203. https://doi.org/10.1093/jpepsy/jst085

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., & Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306–317.

Bollen, K., & Curran, P. (2006). *Latent curve models: A structural equation perspective.* Wiley & Sons, Inc.

Bowers, A., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *Journal of Educational Research, 105*(3), 176–195. https://doi.org/10.1080/00220671.2011.552075

Brandmaier, A. M. (2015). *Semtree: Recursive partitioning of structural equation models in R [Computer software manual]*. http://www.brandmaier.de/semtree

Brandmaier, A., Prindle, J., Lindenberger, U., & McArdle, J. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*(4), 566–582.

Brandmaier, A., Ram, N., Wagner, G., & Gerstorf, D. (2017). Terminal decline in well-being: The role of multi-indicator constellations of physical health and psychosocial correlates. *Developmental psychology, 53*(5), 996–1012. https://doi.org/10.1037/dev0000274

Brandmaier, A., Von Oertzen, T., McArdle, J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71-86.

Brandmaier, A., Von Oertzen, T., McArdle, J., & Lindenberger, U. (2014). Exploratory data mining with structural equation model trees. In J. McArdle & G. Ritschard (Eds.*), Contemporary issues in exploratory data mining in the behavioral sciences* (pp.96-127). Routledge, Taylor and Francis Group.

Breiman, L., & Cutler, A. (2014). *Random forests [Computer software manual]*. http://www.stat.berkeley.edu/breiman/RandomForests/cc_home.html

Choi, J., Elicker, J., Christ, S., & Dobbs-Oates, J. (2016). Predicting growth trajectories in early academic learning: Evidence from growth curve modeling with Head Start children. *Early Childhood Research Quarterly, 36*(1), 244–258. Doi: 10.1016/j.ecresq.2015.12.017

Duprey, M., Pratt, D., Jewell, D., Cominole, M., Fritch, L., Ritchie, E., Rogers, J., Wescott, J., Wilson, D. (2018). *High school longitudinal study of 2009 (HSLS:09): base-year to second follow-up data file documentation (NCES 2018-140).* National Center for Education Statistics. https://nces.ed.gov/pubsearch.

Enders, C., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling, 15*(1), 75-95.

Gottfried, A., Nylund-Gibson, K., Gottfried, A., Morovati, D., & Gonzalez, A. (2017). Trajectories from academic intrinsic motivation to need for cognition and educational attainment. *Journal of Educational Research, 110*(6), 642–652.

Grimm, K., Ram, N., & Estabrook, R. (2017). *Growth modeling: Structural equation and multilevel modeling approaches.* Guilford Press.

Grimm, K., Ram, N., Shiyko, M., & Lo, L. (2014). A simulation study of the ability of growth mixture models to uncover growth heterogeneity. In J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp.150-171). Routledge.

Grubinger, T., Zeileis, A., & Pfeiffer, K. (2011). Evtree: Evolutionary learning of globally optimal classification and regession trees in R. *Journal of Statistical Software, 61*(1), 1-29. https://doi.org/10.18637/jss.v061.i01

Hapfelmeier, A., & Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics and Data Analysis, 60*(1), 50–69.

Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. (2015). Using classification and regression trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and Aging, 30*(4), 911-929.

Hodis, F., Meyer, L., McClure, J., Walkey, F., & Weir, K. (2011). A longitudinal investigation of motivation and secondary school achievement using growth mixture modeling. *Journal of Educational Psychology, 103*(2), 312–323. https://doi.org/10.1037/a0022547

Hu, J., Leite, W., & Gao, M. (2017). An evaluation of the use of covariates to assist in class enumeration in linear growth mixture modeling. *Behavior Research Methods, 49*(3), 1179-1190.

Ingels, S., Pratt, D., Herget, D., Burns, L., Dever, J., Ottem, R., Rogers, J., Jin, Y., Leinwand, S. (2011). *High school longitudinal study of 2009 (HSLS:09): base-year data file documentation (NCES 2011-328).* National Center for Education Statistics. http://nces.ed.gov/pubsearch.

Infurna, F., & Grimm, K. (2017). The use of growth mixture modeling for studying resilience to major life stressors in adulthood and old age: Lessons for class size and identification and model selection. *The Journals of Gerontology: Series B, 73*(1), 148-159.

Jacobucci, R., McArdle, J., & Grimm, K. (2017). A comparison of methods for uncovering sample heterogeneity: Structural equation model trees and finite mixture models. *Structural Equation Modeling, 24(*2), 270–282.

Lee, I., & Rojewski, J. (2013). Brief report: A growth mixture model of occupational aspirations of individuals with high-incidence disabilities. *Journal of Adolescence, 36*(1), 233–239.

Li, L., & Hser, Y. (2011). On inclusion of covariates for class enumeration of growth mixture models. *Multivariate Behavioral Research, 46*(2), 266–302.

Liu, Y., & Lu, Z. (2011). Trajectories of Chinese students' sense of school belonging and academic achievement over the high school transition period. *Learning & Individual Differences, 21*(2), 187–190. https://doi.org/10.1016/j.lindif.2010.12.007

Lubke, G., & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. Structural Equation Modeling, 14(1), 26–47. https://doi.org/10.1207/s15328007sem1401_2

McArdle, J., & Ritschard, G. (2014). *Contemporary issues in exploratory data mining in the behavioral sciences*. Routledge.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

Muthén, B. (2008). Latent variables hybrids: Overview of old and new models. In G. Hancock & K. Samuelsen (Eds.), *latent variable mixture models* (pp.1-24). Information Age Publishing Inc.

Muthén, B. & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.), *Longitudinal data analysis* (pp. 143-165). Taylor & Francis group, LLC.

Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882–891.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*(2), 463-469.

Muthén, L., & Muthén, B. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14(*4), 535–569.

Preacher, K., Wichman, A., MacCallum, R., & Briggs, N. (2008). *Latent growth curve modeling*. Sage Publications, Inc.

R Core Team (2020). *R: A language and environment for statistical computing [Computer software manual]*. http://www.R-project.org/

Ram, N., & Grimm, K. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33(*6), 565–576.

Schumacker, R., & Lomax, R. (2016). *A beginner's guide to structural equation modeling* (4th ed.). Taylor & Francis Groups.

Strobl, C., Tutz, G., & Malley, J. (2009). An Introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14*(4), 323–348.

Tofighi, D. & Enders, C. (2007). Identifying the correct number of classes in growth mixture models. In G. Hancock & K. Samuelsen (Eds.), *latent variable mixture models* (pp.317-341). Information Age Publishing Inc.

Usami, S., Hayes, T., & McArdle, J. (2017). Fitting structural equation model trees and latent growth curve mixture models in longitudinal designs: The influence of model misspecification. *Structural Equation Modeling, 24*(4), 585–598.

Usami, S., Jacobucci, R., & Hayes, T. (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. *Computational Statistics, 34*(1), 1-22.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus* (E-book). Wiley/Higher Education Press. DOI:10.1002/9781118356258