METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Tutorial

Incorporating Machine Learning Into Factor Mixture Modeling: Identification of Covariate Interactions to Explain Population Heterogeneity

Yan Wang [1], Tonghui Xu [2], Jiabin Shen [1]

[1] *Department of Psychology, University of Massachusetts Lowell, Lowell, MA, USA.* [2] *School of Education, University of Massachusetts Lowell, Lowell, MA, USA.*

Methodology, 2023, Vol. 19(3), 303–322, https://doi.org/10.5964/meth.9487

Received: 2022-05-16 • **Accepted:** 2023-06-12 • **Published (VoR):** 2023-09-29

Handling Editor: Katrijn Van Deun, Tilburg University, Tilburg, The Netherlands

Corresponding Author: Yan Wang, Department of Psychology, University of Massachusetts Lowell, 850 Broadway St, Lowell MA 01854, USA. Phone: 978-934-3912. E-mail: Yan_Wang1@uml.edu

Abstract

Factor mixture modeling (FMM) has been widely adopted in health and behavioral sciences to examine unobserved population heterogeneity. Covariates are often included in FMM as predictors of the latent class membership via multinomial logistic regression to help understand the formation and characterization of population heterogeneity. However, interaction effects among covariates have received considerably less attention, which might be attributable to the fact that interaction effects cannot be identified in a straightforward fashion. This study demonstrated the utility of structural equation model or SEM trees as an exploratory method to automatically search for covariate interactions that might explain heterogeneity in FMM. That is, following FMM analyses, SEM trees are conducted to identify covariate interactions. Next, latent class membership is regressed on the covariate interactions as well as all main effects of covariates. This approach was demonstrated using the Traumatic Brain Injury Model System National Database.

Keywords

factor mixture model, latent class, machine learning, structural equation model trees, covariate, interaction

Factor mixture modeling (FMM) has been increasingly used in social, behavioral, and health sciences to examine unobserved population heterogeneity. It enables researchers to model both dimension and typology simultaneously by integrating common factor model and latent class analysis. such that latent classes (i.e., unobserved subgroups) would emerge to capture differences in the common factor model. Latent classes that encapsulate differences in the common factor model among individuals would emerge from the FMM analyses. FMM has been applied with behavioral and health outcomes to examine heterogeneity among psychological trauma victims based on posttraumatic stress disorder symptoms (Elhai et al., 2011), breast cancer patients that reported fatigue symptoms (Ho et al., 2014), and patients with eating disorders based on their emotion regulation profiles (Nordgren et al., 2022), just to list a few.

Among FMM application, covariates (e.g., gender, race) play a critical role in FMM as they are essential to understanding the formation and characterization of latent classes. Specifically, covariates serve as the predictors of latent class membership via multinomial logistic regression in which the log odds of the probability of belonging to a certain class as opposed to a reference class are predicted by covariates. For example, Elhai et al. (2011) found that patients that experienced more traumas and female patients were more likely to be in a more severely symptomatic class as compared with the least symptomatic class.

Despite the prevalence of covariate inclusion, interaction effects among covariates have received considerably less attention. In the context of FMM, covariate interaction refers to the interplay between covariates in affecting latent class membership. In other words, the relationship between latent class membership and one covariate might depend on one or more other covariates. Take children's executive function skills as a hypothetical example. From a developmental perspective, older children have more developed executive function skills compared to their younger counterparts and thus are more likely to be classified into a high executive function class versus a low executive function class. However, this gap in classification between age groups might be smaller for children with severe traumatic brain injuries (TBIs) as executive function skills of both age groups would be negatively affected by the injuries. Therefore, examining covariate interaction effects on latent class membership can offer us a more accurate and nuanced understanding of population heterogeneity, as it is often the complex and multifaceted interplay among factors that impact the outcome. In addition, the identification of covariate interactions can guide the development and implementation of tailored intervention programs that can improve individual outcomes more effectively. For instance, an intervention program to improve the executive function of children with TBIs can leverage the age by TBI severity interaction and tailor its design and/or implementation accordingly.

Although it is critical to identify covariate interactions, they have not been considered or tested in substantive research based on a non-exhaustive review of fifty-nine FMM applications we conducted. Such lack of investigation into covariate interactions in FMM stands in stark contrast to the common testing of interaction effects in other statistical models (e.g., regression) across applied research (Babikian et al., 2011; Ware et al., 2020; Yeates et al., 2010). The lack of attention on covariate interactions in FMM might be attributable to the fact that interaction effects cannot be identified in a straightforward fashion. That is, a major source of covariate selection has been theories or substantive knowledge of researchers; however, it can be a challenging task for applied researchers to come up with hypotheses regarding potential covariate interactions given the unobserved nature of heterogeneity in FMM (Brandmaier et al., 2013; Jacobucci et al., 2017). On the other hand, if an exploratory approach is taken to test all possible interactions, the number of interactions (including higher-order interactions) will increase exponentially as the number of covariates increases, which leads to a complicated model that is difficult to fit and interpret (Moons et al., 2015).

To address this gap in the literature, this study demonstrates the utility of a machine learning approach to identifying covariate interactions that might potentially explain the heterogeneity identified by FMM. Specifically, this study adopted the structural equation model or SEM trees which was proposed by Brandmaier et al. (2013) as a model-based decision tree approach to finding covariates and covariate interactions that impact parameter estimates of the specified model. SEM trees, as other decision tree approaches, have the capacity of automatically searching for covariate interactions (Arnold et al., 2021; Jacobucci et al., 2017). Leveraging this capacity, this study presents a novel integration of SEM trees into FMM for the purpose of identifying potential covariate interactions that explain latent class membership in FMM. This approach was demonstrated using the Traumatic Brain Injury Model System National Database (TBIMS-NDB) April 2020 version), the country's largest multi-center database tracking the rehabilitation trajectories for individuals at least 16 years old treated for inpatient TBI rehabilitation. Through this demonstration, this study aims to provide an exploratory tool for FMM users to identify potential covariate interactions, which offers a more nuanced and sophisticated interpretation of heterogeneity and furthers the understanding of intersectionality.

Factor Mixture Modeling

Factor mixture modeling (FMM) is a combination of common factor model and latent class analysis (LCA), allowing us to model unobserved heterogeneity in parameters of the common factor model. The common factor model can be written as:

$$Y_{ik} = \tau_k + \Lambda_k \eta_{ik} + \varepsilon_{ik}\,. \tag{1}$$

Y_{ik} is a $J \times 1$ vector of responses for an individual i that is assigned to class k (k = 1, 2, ..., K), with J denoting the number of items; τ_k is a $J \times 1$ vector of item intercepts; Λ_k is a $J \times R$ matrix of factor loadings and R refers to the number of factors; η_{ik} is a $R \times 1$ vector of factor scores; and ε_{ik} a $J \times 1$ vector of item residuals that are assumed to be normally distributed with a mean of zero and variance of Θ_k. According to Equation (1), item response is a function of intercepts, factor loadings, factor scores, and residuals, as in a typical common factor model. However, the subscript k associated with the model parameters indicates that they are allowed to vary across latent classes except some constraints needed for model identification. That is, a commonly used identification strategy is to fix the first item loading to be one across classes and the factor mean of the last class is fixed to be zero. Factor scores are assumed to be normally distributed with α_k representing the vector of factor means and $\boldsymbol{\Psi}_k$ the covariance matrix of factors. Thus, the class-specific mean vectors and class-specific variance-covariance matrices can be expressed as:

$$\mu_k = \tau_k + \boldsymbol{\Lambda}_k \alpha_k, \tag{2}$$

$$\Sigma_k = \boldsymbol{\Lambda}_k \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k\,. \tag{3}$$

In FMM, the number of classes is often unknown a priori and needs to be determined by fitting models with varying numbers of classes and comparing model fit using information criteria (ICs), including Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), and sample size adjusted BIC (saBIC; Sclove, 1987). In addition to evaluating model fit, these ICs penalize model complexity by accounting for the number of parameters. Smaller IC values indicate a better trade-off between model fit and model complexity. Additionally, likelihood-based tests can be used in model selection, such as the Lo–Mendell–Rubin test (LMR; Lo et al., 2001), the adjusted LMR (aLMR; Lo et al., 2001), and the bootstrap likelihood ratio test (BLRT; McLachlan & Peel, 2000). These tests compare the fit of models with k and (k-1) classes and a significant test result (e.g., $p < .05$) support the k classes over the (k-1) classes.

In addition to the number of classes, measurement invariance (MI) is an important assumption of valid factor mean comparison across classes that needs to be tested (Clark et al., 2013; Kim et al., 2017; Lubke & Muthén, 2005; Wang et al., 2021). Models with different levels of equality constraints on measurement parameters can be constructed and compared, including configural invariance which requires the same factor structure across classes but factor loadings and intercepts are freely estimated, metric invariance that imposes the equality constraints on factor loadings across classes, and scalar invariance which adds additional equality constraints on intercepts. Note that scalar invariance is often considered as a sufficient prerequisite to factor mean comparison in FMM and multiple-group analyses (Lubke & Muthén, 2005; Meredith, 1993). Beyond MI testing on measurement parameters, the equality of other model parameters (i.e., residual variances, factor variances and covariances) across classes can also be tested to facilitate the understanding and interpretation of latent classes and their differences (Clark et al., 2013).

Structural Equation Model (SEM) Trees

SEM trees integrate SEM into a model-based decision tree paradigm in which the data set is recursively partitioned into subsets based on the splitting of covariates so that differences in SEM parameter estimates are maximized across subsets (Brandmaier et al., 2013; Jacobucci et al., 2017). SEM trees are useful when researchers are interested in finding the influence of covariates and covariate interactions on the SEM model. SEM is a family of statistical procedures that has been widely adopted in social and behavioral sciences to model the relationships among multiple variables (Kline, 2015). One of the key features of SEM is its capacity to model latent constructs (or factors) that are measured by a set of items (or observed variables) and take into account measurement errors. Examples of commonly used SEM procedures include path analysis, the common factor model, structural equation modeling (relationships among multiple factors), and latent growth curve models. Built on the SEM model, SEM trees serve as a tool for exploratory discovery of influences and interactions of covariates on SEM model parameters via the decision tree paradigm.

The decision tree is a supervised machine learning algorithm for prediction and classification (Gupta, 2014; Song & Lu, 2015). It grows a tree structure via recursive partitioning of the covariate space so that individuals classified into the same subset are relatively homogenous in terms of the outcome variable. Figure 1 presents an illustrative example of a scatterplot of a binary outcome variable, diagnosis of the Alzheimer's disease (triangles for Alzheimer's and squares for non-Alzheimer's) on the left and the resultant tree structure on the right, using age and education level as the covariates. The tree structure can be interpreted as a set of “if-then” statements. For instance, if age ≤ 65 and education level ≤ 2, the predicted outcome is Alzheimer’s diagnosis. The splitting of the data set can occur based on multiple criteria and the figure demonstrates a simple rule that constructs a decision tree with a minimal misclassification rate which is also referred to as an incorrect prediction rate (Gupta, 2014).

Algorithms

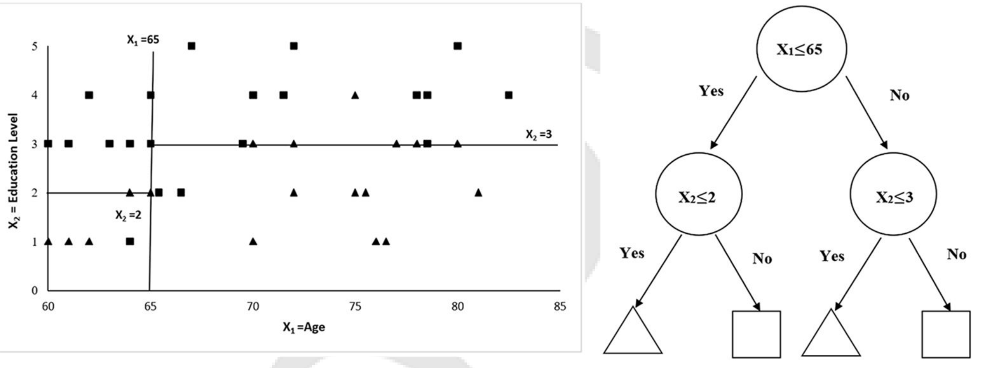
Integrating features of SEM and decision tree, Brandmaier et al. (2013) proposed SEM trees to partition the data set with respect to covariates to maximize difference in SEM parameters across subsets. SEM trees are performed in three steps. First, define a template SEM which is referred to as M, and fit M to the data set. The following equation shows the minimization of a fit function with q degrees of freedom via maximum likelihood estimation (Arnold et al., 2021):

$$F_{ML}\left[\bar{Y}, S, \mu(\theta), \sum(\theta)\right] = \left[\bar{Y} - \mu(\theta)\right]^{T} \sum(\theta)^{-1} \left[\bar{Y} - \mu(\theta)\right] + tr\left[S \sum(\theta)^{-1}\right] - ln\left\{\det\left[S \sum(\theta)^{-1}\right]\right\} - p \tag{4}$$

PsychOpen GOLD

Figure 1

Example of Decision Tree



In this equation, $\bar{Y}$ is a vector of observed means; S is the observed covariance matrix; p indicates the number of observed variables in SEM; θ is a vector of model parameter estimates; $\sum(\theta)$ is the model-implied covariance matrix; and $\mu(\theta)$ is a vector of model-implied means.

Second, to evaluate a possible split based on a covariate, the full data is partitioned into l subsets where $l = 1, \quad 2, \quad \ldots, L$, and the template SEM model is fitted to each subset. Given that the subsets are non-overlapping, the fit of all SEMs across subsets is evaluated independently based on Equation (4) and these models are referred to as M_{SUB}. Then the fit of M_{SUB} and M is compared using the likelihood ratio test:

$$LR = (N-1)\left\{F_{ML}\left[\bar{Y}_F, S_F, \mu\left(\hat{\theta}_F\right), \Sigma\left(\hat{\theta}_F\right)\right] - \sum_{l=1}^{L} \frac{n_l}{N} F_{ML}\left[\bar{Y}_l, S_l, \mu\left(\hat{\theta}_l\right), \Sigma\left(\hat{\theta}_l\right)\right]\right\} \tag{5}$$

N and n_l refer to the sample size for the full data set and the subset l. LR follows the chi-square distribution with $(L-1)q$ degrees of freedom. All possible splits are evaluated for each covariate, and the split with maximum increase in the LR is chosen.

Lastly, repeat the steps for each subset due to the chosen split to find further partitions that significantly improve the model fit; if the partition does not improve the model fit, then further partitioning is terminated. Results of SEM trees can be visualized as a tree structure with nodes. The inner node (i.e., node that has successors) represents a cut point with respect to a covariate, and leaf nodes are associated with an SEM that represents the induced subsamples of the data (Brandmaier et al., 2013).

Model Constraints

Similar to FMM, constraints on SEM model parameters can be imposed in SEM trees. Specifically, there are two types of restrictions in a tree: a global restriction and a local restriction. A global restriction can be imposed on any parameter(s) in the SEM model in which the value for the constrained parameter is estimated with the full data set and fixed across all subsequent models. A local restriction is imposed only for split evaluation such that the parameters are equal across all models that share the same inner node, but the resultant leaf nodes can have different values of the parameters. In other words, parameters are allowed to be different across models, but their differences do not contribute to the split evaluation.

Integrating SEM Trees Into FMM

Among a few applications of SEM trees that have been identified (Ammerman et al., 2019; de Mooij et al., 2018; Li et al., 2021; Sagan & Łapczyński, 2020), interaction among covariates was present. For instance, Li et al. (2021) included a total of 33 covariates to examine their associations with students' attitudes towards collaboration, and found that student gender affected the CFA model parameters of students' attitudes towards collaboration, but only for those with above-average home educational resources, which indicated an interaction effect between gender and home educational resources. Given the advantage of SEM trees in automatically searching for covariate interactions, this study proposes an integrated use of SEM trees and FMM such that covariate interactions that are identified by SEM trees might potentially explain heterogeneity in FMM.

The proposed integrated use consists of the following five steps:

1. Identify constructs and items for the FMM analyses, as well as covariates that might potentially explain the distinction among latent classes. Constructs refer to the latent factors that are measured by a set of items, which is the basis of FMM analyses as shown in Equation (1).
2. Conduct unconditional FMM analyses (without covariates) based on the identified constructs and items. Specifically, given that the number of classes and the class-varying parameters are unknown, a series of FMMs can be specified and fitted to the data, including 1-class, 2-class configural, metric, and scalar invariance models, 3-class configural, metric, and scalar invariance models, etc. The fitted models can be compared in terms of fit based on multiple ICs, such as AIC, BIC, and saBIC[1]. Model with the smallest ICs can be chosen as the best-fitting model.

1) LMR, aLMR, and BLRT were not used because they are appropriate for determining the number of classes; however, compared models in the analysis involves different class-varying parameters in addition to the number of classes. Thus, the likelihood-based tests were not appropriate.

https://doi.org/10.5964/meth.9487

3. Examine the substantive interpretability of the best-fitting model based on parameter estimates.
4. Conduct SEM trees analyses to identify covariate interactions that could potentially explain latent class membership in FMM. To maximize the chance that covariate interactions selected by the SEM trees would explain latent class membership in FMM, we propose that the specification of parameter restrictions between these two approaches should be matched. That is, the level of invariance (i.e., configural, metric, or scalar) that is identified in FMM is also adopted in SEM trees via the global constraint function.
5. Multinomial logistic regression is conducted with covariate interactions that are detected by the SEM trees as well as all main effects to examine correlates of latent classes. The three-step approach to covariate inclusion is adopted here, given that the identification of latent classes is done without the influence of covariates, and the impact of covariates and covariate interactions is examined while taking into account classification errors (Asparouhov & Muthén, 2014; Vermunt, 2010).

Demonstration

This demonstration serves as example of the integrated use of FMM and SEM trees via the five steps proposed above. The sample came from the Traumatic Brain Injury Model System National Database (TBIMS-NDB) obtained as public datasets with version date of April 2020. TBIMS-NDB was funded by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR) as a prospective, longitudinal, multicenter database to examine the health outcomes of more than 17,000 individuals who experienced TBIs that require inpatient rehabilitation in the United States. All data were collected using surveys, with baseline data collected at the time of discharge from inpatient rehabilitation settings and follow-up data collected at 1-, 2-, 5-, 10-, 15-, 20-, 25-, and 30-years post-injury. This demonstration used the 1-year post-injury data that consisted of 9,741 individuals. A full description of the sociodemographic characteristics of the sample as well as other descriptive statistics of the variables is provided in Table 1. Annotated codes for the following analyses are included in the electronic Supplementary Materials.

For Step 1, the 5-item Satisfaction with Life Scale (SWLS) was used as the outcome assessment for life satisfaction levels among individuals following TBI (Diener et al., 1985; Pavot & Diener, 1993). Each item scored from 1 (lowest life satisfaction) to 7 (highest life satisfaction) asking different aspects of a patient's perception of his/her life conditions. A total of seven covariates were identified, including Functional Independence Measure (FIM) Cognitive on Admission (Linacre et al., 1994), pre-injury disability and pre-injury limitations (National Research Council, 2004), TBI severity (Teasdale & Jennett, 1976) as measured by patients' total Glasgow Coma Scores, age at injury, biological sex, race,

Table 1

Descriptive Statistics of Variables and Sample Sociodemographic Characteristics

Variable/Characteristic	Statistic		
Life Satisfaction	***N***	***M***	***SD***
1. Ideal life	9717	4.06	2.08
2. Excellent life conditions	9728	4.06	2.08
3. Satisfaction with life	9729	4.60	2.05
4. Important things in life	9723	4.71	1.99
5. Life lived over	9709	3.84	2.22
Continuous Covariates	***N***	***M***	***SD***
TBI severity	5529	11.21	4.06
FIM Cognition	9695	16.03	7.58
Categorical Covariates	***N***	***%***	
Sex			
Females	2751	28.25	
Males	6988	71.75	
Race			
White	6897	70.82	
Black	1596	16.39	
Hispanic	849	8.72	
Others	397	4.08	
Age Group			
AYAs	2994	30.74	
Adults	5108	52.44	
Older Adults	1639	16.83	
Pre-Injury Employment Status			
Employed	6389	66.12	
Student	706	7.31	
Unemployed	2568	26.58	
Pre-Injury Impairment			
Yes	368	5.49	
No	6333	94.51	
Pre-Injury Physical Limitation			
Yes	491	7.33	
No	6206	92.67	

Note. Ideal life = In most ways my life is close to my ideal; Excellent life conditions = The conditions of my life are excellent; Satisfaction with life = I am satisfied with my life; Important things in life = I have gotten important things I want in life; Life lived over = If I could live my life over, I would change almost nothing. AYAs = adolescents and young adults.

and pre-injury employment status. All covariates were collected at baseline visit. Age at injury was recoded as a categorical variable: adolescents and young adults (AYAs; ≤ 25), adults (26–59), and older adults or seniors (≥ 60).

For Step 2, unconditional FMM analyses were conducted with life satisfaction in M*plus* 8.4[2] (Muthén & Muthén, 1998-2017). Table 2 presents model fit comparisons of FMMs. All fitted models converged except the 4-class configural and scalar models. Among converged models, AIC, BIC, and saBIC consistently showed that the 4-class metric model had a superior fit.

Table 2

Model Fit Comparison of Factor Mixture Modeling

Model	Parm	LL	AIC	BIC	saBIC	Entropy	Class Proportions
1-class	15	-94483	188996	189104	189056		
2-class conf	31	-88689	177440	177663	177565	.90	.72/.28
2-class metric	27	-88795	177644	177838	177753	.90	.73/.27
2-class scalar	18	-93401	186838	186967	186910	.92	.38/.62
3-class conf	47	-85263	170619	170957	170807	.91	.14/.58/.28
3-class metric	39	-85345	170769	171049	170925	.91	.14/.58/.28
3-class scalar	21	-93411	186863	187014	186947	.65	.40/.39/.21
4-class conf				Non-convergence			
4-class metric	51	-84430	168961	169328	169166	.87	.14/.25/.33/.28
4-class scalar				Non-convergence			

Note. conf = configural invariance; metric = metric invariance; scalar = scalar invariance; Parm = number of free parameters; LL = log-likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion; saBIC = sample size adjusted BIC.

For Step 3, interpretability of the 4-class metric model was examined. Table 3 presents the parameter estimates of this model by latent class. While loadings were constrained to be equal across classes, intercepts, factor mean, and factor variance were allowed to be freely estimated.[3] Factor means were estimated to be -4.61, -3.01, and -1.98 for Classes 1, 2, and 3 respectively, with Class 4 serving as the reference group (factor mean 0). Note that although factor mean comparison is not permitted with a metric invariance model, factor means of Classes 1, 2, and 3 were statistically significantly different from zero. Class 3 had the largest proportion, .33, followed by Class 4 (.28), Class 2 (.25), and Class 1 (.14).

2) The EM algorithm was used to find the optimal parameter estimates via an iterative process until the convergence criterion (.00005 by default of M*plus*) was met.

3) Exceptions were that intercept of the first item was constrained to be equal across classes and the factor mean of the last class (i.e., Class 4) in M*plus* was fixed to be zero, for the identification purpose.

Table 3

Parameter Estimates of the Four-Class Metric Invariance FMM

		Intercept			
Item/Statistic	**Loading**	**Class 1**	**Class 2**	**Class 3**	**Class 4**
Item					
Ideal	1.00	6.12	6.12	6.12	6.12
Cond	1.15	6.87	6.47	6.48	6.13
Satisfied	1.05	6.45	5.78	8.00	6.24
Important	.94	6.90	6.79	6.74	6.21
Live again	.88	5.57	6.16	5.64	5.23
Statistic					
Factor mean		-4.61	-3.01	-1.98	0
Factor variance		.23	.43	.34	.32
Class proportion		.14	.25	.33	.28

Distinction of the latent classes was further interpreted based on the life satisfaction item mean by class, as illustrated in Figure 2. ANOVAs with Bonferroni adjustment were conducted to compare the item means across classes and results showed statistically significant mean differences between any two groups. Class 4 had the highest mean across all items, followed by Class 3, Class 2, and Class 1. Of note is that Class 3 had relatively high mean on the item, "I am satisfied with my life", which might correspond to the high item intercept in the 4-class metric invariance FMM.

For Step 4, SEM trees were performed in the *semtree* package in R (Brandmaier et al., 2021; R Core Team, 2021). A CFA model of life satisfaction measured by five items was specified and a total of 12 covariates were included. Given that a 4-class metric invariance model was supported in FMM, metric invariance was also established in SEM trees via the global constraints function such that factor structures and loadings were constrained to be equal across groups whereas intercepts, factor mean, and residual variances were freely estimated. The resulting tree was displayed in Figure 3. There were four splits among which the first two occurred on age and the other two on race. The first split divided the whole sample into two, older adults (n = 1639) versus the rest (n = 8102). The second split further divided those that were not older adults into two, adults (n = 5108) versus AYAs (n = 2994). Each of these two groups was split again on whether or not the patient was Black. Therefore, there were a total of five groups as a result of SEM trees, older adults, Black adults, adults that were not Black, Black AYAs, and AYAs that were not Black, n = 1639, 921, 4187, 502, 2490 respectively.

Figure 2

Life Satisfaction Item Mean by Latent Class

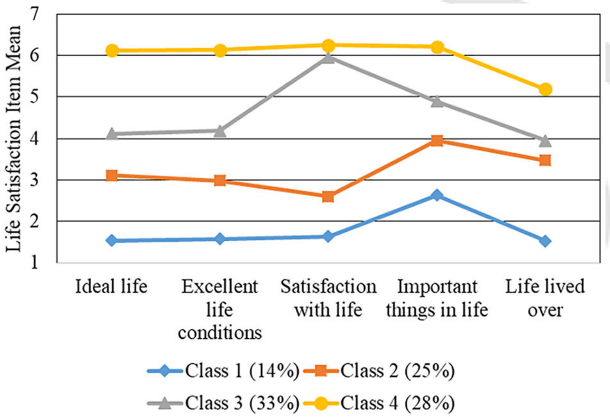
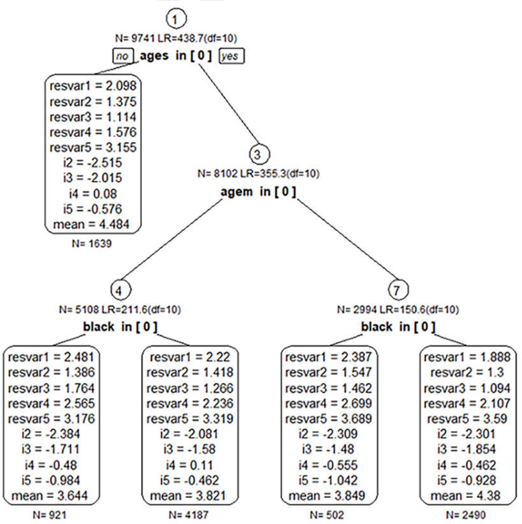


Figure 3

Tree Plot of SEM Trees



Note. N refers to the sample size at each split; LR is the likelihood ratio statistic with the difference in degrees of freedom (*df*); ages and agem refer to older adults and adults, respectively; black refers to the race group of Black.

Table 4

Results of Multinomial Logistic Regression via the Three-Step Approach

	Class 1		Class 2		Class 3	
Covariate	**Est (*SE*)**	**OR**	**Est (*SE*)**	**OR**	**Est (*SE*)**	**OR**
TBI severity	-.04 (.02)	0.96*	-.01 (.01)	0.99	-.01 (.01)	0.99
FIM cognition	-.01 (.01)	0.99	-.02 (.01)	0.98*	-.01 (.01)	0.99
Adults	.63 (.18)	1.87***	.51 (.14)	1.66***	-.21 (.13)	0.81
Older Adults	-.56 (.24)	0.57*	-.06 (.18)	0.94	-.63 (.16)	0.54***
Female	.04 (.14)	1.04	.12 (.11)	1.12	.10 (.11)	1.10
Black	.72 (.18)	2.06***	.81 (.16)	2.24***	.54 (.16)	1.71**
Hispanic	.05 (.20)	1.05	.22 (.16)	1.24	-.10 (.16)	0.90
OtherRace	-.58 (.39)	0.56	.37 (.22)	1.44	-.28 (.24)	0.76
Student	-.10 (.33)	0.91	.07 (.24)	1.07	.04 (.22)	1.04
Unemployed	.64 (.15)	1.89***	.28 (.12)	1.32*	.29 (.11)	1.34**
Pre-impairment	-.22 (.27)	0.80	-.002 (.20)	1.00	.02 (.19)	1.02
Pre-phylimit	.38 (.22)	1.47	.16 (.18)	1.18	.18 (.18)	1.19
Older Adults*Black	-.82 (.52)	0.44	-.88 (.35)	0.42*	-.29 (.32)	0.75

Note. Pre-impairment = pre-injury impairment; pre-phylimit = pre-injury physical limitation; the missing groups for categorical covariates are the reference groups (i.e., AYAs, Male, White, and Employed). Est (*SE*) = estimated regression coefficient (standard error); OR = odds ratio.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Given that split occurred on whether or not the patient was Black for both adults and AYAs but not older adults, an interaction effect was signified between the race category of Black and older adults. In other words, the impact of being Black on CFA model parameters was absent for older adults and present for the rest of the sample.

For Step 5, the interaction effect between older adults and Black that was detected by SEM trees was included in the multinomial logistic regression on top of all main effects. Results (see Table 4) showed that the interaction effect was significant for Class 2, B(*SE*) = -.88(.35), p = .013, which indicates that the impact of race on the likelihood of being assigned to Class 2, a somewhat satisfaction class, depended upon age group. That is, for individuals that were AYAs, the odds of being in Class 2 (versus Class 4, the reference group) for Black people were 2.24 times that of White people, controlling for all other covariates in the model. However, for older adults, Black individuals experienced a reduction of 7% in the odds of being in Class 2 compared to the White. In other words, seniority positively related with life satisfaction for Black individuals, and the Black AYAs were at a higher risk for life dissatisfaction.

The interaction between age group and race is further illustrated in Table 5 in which the composition of Classes 2 and 4 with regards to age group and race is presented.

Table 5

Age Group by Race Interaction Effect

Race and Age Group	Class 2	Class 4
Black		
AYAs	119 (27.36%)	80 (29.52%)
Adults	283 (65.06%)	135 (49.82)
Older Adults	33 (7.59%)	56 (20.66%)
Total	435 (100.00%)	271 (100.00%)
White		
AYAs	378 (23.46%)	664 (31.77%)
Adults	929 (57.67%)	926 (44.31%)
Older Adults	304 (18.87%)	500 (23.92%)
Total	1611 (100.00%)	2090 (100.00%)

Note. AYAs = adolescents and young adults.

That is, among 435 Black people that were assigned to Class 2, the somewhat satisfaction class, only 7.59% were senior, whereas 20.66% of Black people in Class 4, the high satisfaction class, were senior. The discrepancy in percentages was not as substantial as above for the Black AYAs, the White seniors, or the White AYAs. In addition to the interaction effect, adults were more likely to be in Class 2 than AYAs and those that were unemployed were associated with a higher likelihood of being in Class 2 than those that were employed.

For the other classes (i.e., Classes 1 and 3), despite the absence of a significant interaction effect, age, race, and unemployment all had significant impact on the latent class membership. That is, adults were more likely to be in Class 1 which were characterized by low life satisfaction, compared with AYAs. Older adults were less likely to be in Classes 1 and 3 which were the low and moderate life satisfaction classes, respectively, compared with AYAs. Individuals who were Black were more likely to be in Classes 1 and 3 than Class 4, compared with those that were White. Those that were unemployed were associated with a higher likelihood of being in Classes 1 and 3 compared with those that were employed.

Discussion

This study aimed to demonstrate the utility of a machine learning approach, SEM trees, for the identification of covariate interactions that potentially explain latent classes in FMM. Specifically, this study tapped into the advantage of SEM trees in automatically searching for covariate interactions and showed that covariate interaction that was detected by SEM trees can be incorporated into FMM to explain the distinction among latent classes. As demonstrated, SEM trees revealed the interaction between race and age group, which provided a more nuanced understanding of how these factors interplayed to affect life satisfaction. That is, the impact of being Black on individuals' likelihood of being assigned to a somewhat satisfaction versus a high satisfaction class depended on age group, which clearly indicates seniority as a protective factor against life dissatisfaction. Retrospectively, this interaction effect is in alignment with the prior literature on life satisfaction and other psychological and health outcomes (Ajrouch et al., 2001; George et al., 1985; Phatak et al., 2013; Shaw et al., 2010). Overall, this demonstration provides an example of how intersectionality can be examined and understood with an integration of FMM and SEM trees.

Despite the utility of the SEM trees in identifying covariate interactions, there is no guarantee that the interaction terms will turn out to be the sources of heterogeneity in FMM. For example, the race by age group interaction was statistically significant in one latent class, but not for the other two classes. This possible discrepancy between FMM and SEM Trees occurred due to the drastic differences between the two approaches in how heterogeneity is modeled (Jacobucci et al., 2017). That is, in FMM, latent classes formed on the basis of the estimated model parameters (e.g., intercepts, loadings, factor mean, factor variance), whereas splits of the sample in SEM trees depend upon covariates. Note that although a conditional FMM might be more comparable to SEM trees given that the contribution of covariates to the formation of latent classes is allowed, we adopted unconditional FMM in our study which allows researchers to first examine heterogeneity based on the outcome of interest and subsequently explore the impact of covariates. This has been aligned with the vast majority of FMM applications (e.g., Babusa et al., 2015; Bernstein et al., 2013; Elhai et al., 2011).

The possible discrepancy between FMM and SEM trees in identifying covariate interactions does not undermine the utility of SEM trees in suggesting potential interactions. Especially when intersectionality is of interest to applied researchers but substantive theories or knowledge regarding the form of interactions are lacking, SEM trees offers a data-driven and exploratory approach that can be adopted to identify possible interaction effects that explain latent classes in FMM. As demonstrated in the paper, an unconditional FMM can be conducted first to identify latent classes and the level of equality constraints on parameters across classes. Next, the SEM trees can be conducted with a comparable level of constraints to FMM (e.g., loadings are equal across classes) and the suggested covariate interactions could be added to the multinomial logistic regression on top of the main effects via the three-step approach. Alternatively, if hypothesis regarding interaction effects is available, the two modeling approaches can be used concurrently and SEM trees at least offer an alternative perspective into how heterogeneity is shaped by covariates.

While we highlight the utility of SEM trees in suggesting covariate interactions, a few caveats are worth mentioning. First, future Monte Carlo simulation studies are needed to systematically evaluate the efficacy of this approach of integrating SEM trees with FMM. For example, multiple splitting methods and options to control the growth of the tree are available in the implementation of the SEM trees approach, and simulation studies are needed to examine which method and option would be optimal under which data conditions (Jacobucci et al., 2017). Additional factors that can be considered in simulation studies include numbers of latent classes, degrees of class separation, number of covariates, forms of interactions (e.g., two-way or higher-order interactions), etc. Second, the SEM trees approach should not be considered as a replacement of substantive theories or knowledge in identifying covariate interactions (Brandmaier et al., 2013). Covariate interactions suggested by the SEM trees should be meaningful and interpretable through a retrospective check with theories or knowledge of researchers, prior to the addition of interactions into the multinomial logistic regression. Third, this study demonstrated the utility of the SEM trees for FMM and future research is needed to examine the potential of this approach for other mixture models (e.g., growth mixture model, latent class analysis) via demonstrations and Monte Carlo simulations. Despite these caveats, we encourage FMM users to tap into the advantage of the SEM trees in identifying potential covariate interactions that advance their understanding of intersectionality and heterogeneity.

Funding: This research was supported by the American Educational Research Association Division D (000000000035095); the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (R00HD093814). The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Educational Research Association Division D or the National Institutes of Health.

Acknowledgments: The Traumatic Brain Injury (TBI) Model Systems National Database is a multicenter study of the TBI Model Systems Centers Program, and is supported by the National Institute on Disability, Independent Living and Rehabilitation Research (NIDILRR), a center within the Administration for Community Living (ACL), Department of Health and Human Services (HHS). However, these contents do not necessarily reflect the opinions or views of the TBI Model Systems Centers, NIDILRR, ACL or HHS.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: The sample data for the demonstration above can be requested by researchers from the National Data and Statistical Center, NDSC (https://www.tbindsc.org).

Supplementary Materials

The supplementary materials provided are the annotated codes for unconditional FMM analyses, annotated codes for SEM Trees, and the annotated codes for the three-step approach to estimate covariate and covariate interaction effect on latent class membership (see Wang et al., 2023).

Index of Supplementary Materials

Wang, Y., Xu, T., & Shen, J. (2023). *Supplementary materials to "Incorporating machine learning into factor mixture modeling: Identification of covariate interactions to explain population heterogeneity"* [Model systems program, Annotated codes]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.13269

References

Ajrouch, K. J., Antonucci, T. C., & Janevic, M. R. (2001). Social networks among Blacks and Whites: The interaction between race and age. *Journals of Gerontology: Series B, Psychological Sciences and Social Sciences, 56*(2), S112–S118. https://doi.org/10.1093/geronb/56.2.S112

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ammerman, B. A., Jacobucci, R., & McCloskey, M. S. (2019). Reconsidering important outcomes of the nonsuicidal self-injury disorder diagnostic criterion A. *Journal of Clinical Psychology, 75*(6), 1084–1097. https://doi.org/10.1002/jclp.22754

Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology, 11*, Article 564403. https://doi.org/10.3389/fpsyg.2020.564403

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using M*plus*. *Structural Equation Modeling, 21*(3), 329–341. https://doi.org/10.1080/10705511.2014.915181

Babikian, T., Satz, P., Zaucha, K., Light, R., Lewis, R., & Asarnow, R. (2011). The UCLA longitudinal study of neurocognitive outcomes following mild pediatric traumatic brain injury. *Journal of the International Neuropsychological Society, 17*(5), 886–895. https://doi.org/10.1017/S1355617711000907

Babusa, B., Czeglédi, E., Túry, F., Mayville, S. B., & Urbán, R. (2015). Differentiating the levels of risk for muscle dysmorphia among Hungarian male weightlifters: A factor mixture modeling approach. *Body Image, 12*, 14–21. https://doi.org/10.1016/j.bodyim.2014.09.001

Bernstein, A., Stickle, T. R., & Schmidt, N. B. (2013). Factor mixture model of anxiety sensitivity and anxiety psychopathology vulnerability. *Journal of Affective Disorders, 149*(1–3), 406–417. https://doi.org/10.1016/j.jad.2012.11.024

Brandmaier, A. M., Prindle, J. J., & Arnold, M. (2021). *Recursive partitioning for structural equation model trees* [Computer software manual]. R Foundation for Statistical Computing. https://cran.r-project.org/web/packages/semtree/semtree.pdf

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71–86. https://doi.org/10.1037/a0030001

Clark, S. L., Muthén, B. O., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structural underlying psychological disorders. *Structural Equation Modeling, 20*(4), 681–703. https://doi.org/10.1080/10705511.2013.824786

de Mooij, S. M. M., Henson, R. N. A., Waldorp, L. J., & Kievit, R. A. (2018). Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience, 38*(25), 5826–5836. https://doi.org/10.1523/JNEUROSCI.1627-17.2018

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Elhai, J. D., Naifeh, J. A., Forbes, D., Ractliffe, K. C., & Tamburrino, M. (2011). Heterogeneity in clinical presentations of posttraumatic stress disorder among medical patients: Testing factor structure variation using factor mixture modeling. *Journal of Traumatic Stress, 24*(4), 435–443. https://doi.org/10.1002/jts.20653

George, L. K., Okun, M. A., & Landerman, R. (1985). Age as a moderator of the determinants of life satisfaction. *Research on Aging, 7*(2), 209–233. https://doi.org/10.1177/0164027585007002004

Gupta, G. K. (2014). *Introduction to data mining with case studies.* PHI Learning.

Ho, R. T. H., Fong, T. C. T., & Cheung, I. K. M. (2014). Cancer-related fatigue in breast cancer patients: Factor mixture models with continuous non-normal distributions. *Quality of Life Research, 23*(10), 2909–2916. https://doi.org/10.1007/s11136-014-0731-7

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2017). A comparison of methods for uncovering sample heterogeneity: Structural equation model trees and finite mixture models. *Structural Equation Modeling, 24*(2), 270–282. https://doi.org/10.1080/10705511.2016.1250637

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling, 24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kline, R. B. (2015). *Principles and practice of structural equation modeling.* Guilford Publications.

Li, J., Zhang, M., Li, Y., Huang, F., & Shao, W. (2021). Predicting students' attitudes toward collaboration: Evidence from Structural Equation Model Trees and Forests. *Frontiers in Psychology, 12*, Article 604291. https://doi.org/10.3389/fpsyg.2021.604291

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75*(2), 127–132. https://doi.org/10.1016/0003-9993(94)90384-0

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. https://doi.org/10.1007/BF02294825

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine, 162*(1), W1–W73. https://doi.org/10.7326/M14-0698

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nordgren, L., Ghaderi, A., Ljótsson, B., & Hesser, H. (2022). Identifying subgroups of patients with eating disorders based on emotion dysregulation profiles: A factor mixture modeling approach to classification. *Psychological Assessment, 34*(4), 367–378. https://doi.org/10.1037/pas0001103

National Research Council. (2004). *The 2000 census: Counting under adversity.* National Academies Press.

Pavot, W., & Diener, E. (1993). Review of the Satisfaction with Life Scale. *Psychological Assessment, 5*(2), 164–172. https://doi.org/10.1037/1040-3590.5.2.164

Phatak, U. R., Kao, L. S., Millas, S. G., Wiatrek, R. L., Ko, T. C., & Wray, C. J. (2013). Interaction between age and race alters predicted survival in colorectal cancer. *Annals of Surgical Oncology, 20*(11), 3363–3369. https://doi.org/10.1245/s10434-013-3045-z

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software manual]. R Foundation for Statistical Computing. https://www.R-project.org/

Sagan, A., & Łapczyński, M. (2020). SEM-Tree hybrid models in the preference analysis of the members of Polish households. *Advances in Data Analysis and Classification, 14*(4), 855–869. https://doi.org/10.1007/s11634-020-00414-7

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360

Shaw, B. A., Liang, J., & Krause, N. (2010). Age and race differences in the trajectories of self-esteem. *Psychology and Aging, 25*(1), 84–94. https://doi.org/10.1037/a0018242

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Teasdale, G., & Jennett, B. (1976). Assessment and prognosis of coma after head injury. *Acta Neurochirurgica, 34*(1–4), 45–55. https://doi.org/10.1007/BF01405862

Traumatic Brain Injury Model Systems Program. (2023). *Traumatic Brain Injury Model System National Database (TBIMS-NDB)* (Version April 2020) [Data set]. Traumatic Brain Injury Model Systems National Data and Statistical Center. https://www.tbindsc.org/

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Wang, Y., Kim, E., Ferron, J. M., Dedrick, R. F., Tan, T. X., & Stark, S. (2021). Testing measurement invariance across unobserved groups: The role of covariates in factor mixture modeling.

Educational and Psychological Measurement, 81(1), 61–89. https://doi.org/10.1177/0013164420925122

Ware, A. L., Shukla, A., Goodrich-Hunsaker, N. J., Lebel, C., Wilde, E. A., Abildskov, T. J., Bigler, E. D., Cohen, D. M., Mihalov, L. K., Bacevice, A., Bangert, B. A., Taylor, H. G., & Yeates, K. O. (2020). Post-acute white matter microstructure predicts post-acute and chronic post-concussive symptom severity following mild traumatic brain injury in children. *NeuroImage: Clinical, 25*, Article 102106. https://doi.org/10.1016/j.nicl.2019.102106

Yeates, K. O., Taylor, H. G., Walz, N. C., Stancin, T., & Wade, S. L. (2010). The family environment as a moderator of psychosocial outcomes following traumatic brain injury in young children. *Neuropsychology, 24*(3), 345–356. https://doi.org/10.1037/a0018387

European Association of Methodology

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.