frontiers in Education

ORIGINAL RESEARCH
published: 04 March 2021
doi: 10.3389/feduc.2021.613645

Model Fit and Comparison in Finite Mixture Models: A Review and a Novel Approach

Kevin J. Grimm, Russell Houpt and Danielle Rodgers*

Department of Psychology Arizona State University, Tempe, AZ, United States

One of the greatest challenges in the application of finite mixture models is model comparison. A variety of statistical fit indices exist, including information criteria, approximate likelihood ratio tests, and resampling techniques; however, none of these indices describe the amount of improvement in model fit when a latent class is added to the model. We review these model fit statistics and propose a novel approach, the *likelihood increment percentage per parameter* (*LIPpp*), targeting the relative improvement in model fit when a class is added to the model. Simulation work based on two previous simulation studies highlighted the potential for the *LIPpp* to identify the correct number of classes, and provide context for the magnitude of improvement in model fit. We conclude with recommendations and future research directions.

Keywords: latent class analyses, growth mixture modeling, model comparison, finite mixture model, latent profile analysis

OPEN ACCESS

Edited by:
Katerina M. Marcoulides, University of Minnesota Twin Cities, United States

Reviewed by:
Ren Liu, University of California, Merced, United States
Jam Khojasteh, Oklahoma State University, United States

****Correspondence:***
Kevin J. Grimm
kjgrimm@asu.edu

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *02 October 2020*
Accepted: *20 January 2021*
Published: *04 March 2021*

Citation:
Grimm KJ, Houpt R and Rodgers D (2021) Model Fit and Comparison in Finite Mixture Models: A Review and a Novel Approach. Front. Educ. 6:613645. doi: 10.3389/feduc.2021.613645

INTRODUCTION

Finite mixture modeling (FMM) is a broad class of statistical models to examine whether model parameters vary over unmeasured groups of individuals. Researchers have fit FMMs to search for unmeasured groups in regression analysis (e.g., Liu and Lin, 2014), factor models (e.g., Lubke and Muthén, 2005), growth models (Muthén and Shedden, 1999), mixed-effects models (Wang et al., 2002), and FMMs are the foundation of latent class analysis (Lazarsfeld, 1950) and latent profile analysis (Gibson, 1959). When applied to empirical data, the number of unmeasured groups, often referred to as *latent classes*, and the model parameters that differ over the unmeasured groups are unknown. Thus, a series of FMMs, differing in the number of latent classes and the nature of model constraints, are specified, fit, and then compared. The comparison of FMMs is not straightforward because models with different constraints are not necessarily nested. Moreover, FMMs that differ in the number of classes, but have the same parameter constraints are nested; however, the difference in −2 log-likelihood is not chi-square distributed under the null hypothesis. Because of these issues, different model fit criteria are used for model selection. In this paper, we review model fit criteria to compare FMMs and Monte Carlo simulation studies that have investigated the performance of these model comparison approaches. We then propose a novel approach to model comparison for FMMs based on the relative improvement in model fit.

Model Comparison Criteria

Model comparison criteria for FMMs fall into one of three categories. The first category is *information criteria*, such as the Bayesian Information Criterion (*BIC*; Schwarz, 1978); the second category is *approximate likelihood ratio tests*, such as the *Bootstrap Likelihood Ratio Test* (BLRT; McLachlan,

1987); and the third category relies on resampling techniques, such as the *k*-fold cross validation approach (Grimm et al., 2017).

Information Criteria

Information criteria combine the −2 log-likelihood (−2*LL*) from the model, which is a measure of how well the model fits the data, and a penalty for each model parameter. With multivariate Gaussian data, the −2*LL* is

$$-2 \cdot \sum_{i=1}^{N}\left(-\frac{k_i}{2}\ln(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(y_i-\mu_i)'\Sigma_i^{-1}(y_i-\mu_i)\right) \quad (1)$$

where k_i is the number of observations for person *i*, y_i is the vector of observed scores for person *i*, Σ_i is the model implied covariance matrix based on person *i*'s measured variables, and μ_i is the model implied mean vector based on person *i*'s measured variables. The elements of the −2*LL* are subscripted by *i* because their dimensions depend on the number of measured variables for person *i*. The −2*LL* is a measure of misfit with higher values indicative of a greater discrepancy between the observed data and the model-implied mean vector and covariance matrix.

Information criteria take on the general form

$$-2LL + p \cdot penalty \quad (2)$$

where *p* is the number of estimated parameters and *penalty* is the amount the model's fit is penalized for each estimated parameter. The penalty for each parameter is often a simple function and may depend on sample size. Because information criteria have a penalty for the number of estimated parameters, the model with the lowest information criteria is typically selected.

There are many information criteria that differ in the penalty term. Commonly reported information criteria include the Akaike Information Criterion (*AIC*; Akaike, 1973), the *BIC* (Schwarz, 1977), the sample size adjusted *BIC* (*saBIC*; Sclove, 1987), and the corrected *AIC* (*AICc*; Hurvich and Tsai, 1989). Information criteria are used to compare both nested and non-nested models, which makes them appropriate for comparing FMMs with a different number of classes and/or different parameter constraints across classes.

The *AIC* is defined as

$$AIC = -2LL + 2p \quad (3)$$

where the penalty for each parameter is 2. The *AIC* has a constant penalty for each parameter, which is unique for an information criterion. The *AICc* was proposed to improve performance of the *AIC* in small samples (Hurvich and Tsai, 1989), where the *AIC* was found to prefer overparameterized models. The *AICc* applies an adjustment for sample size and is written as

$$AICc = AIC + \frac{2p^2+2p}{N-p-1} = -2LL + 2p + \frac{2p^2+2p}{N-p-1} \quad (4)$$

where *p* is the number of estimated parameters and *N* is the sample size. As sample size increases, the adjustment (i.e., $\frac{2p^2+2p}{N-p-1}$) converges toward 0 making the difference between the *AICc* and the *AIC* negligible in large samples.

The *BIC* is

$$BIC = -2LL + \ln(N) \cdot p \quad (5)$$

where $\ln(N)$ is the natural log of the sample size and the penalty for each estimated parameter. Compared to the *AIC*, the *BIC*'s penalty for each parameter, $\ln(N)$, is larger when $N \geq 8$. Thus, the *BIC* generally favros models with fewer parameters compared to the *AIC*. In 1987, Sclove proposed the *saBIC*, which replaces *N* in **Eq. 5** with $\left(\frac{N+2}{24}\right)$. Compared to the *BIC*, the per parameter penalty in the *saBIC* is smaller making the *saBIC* favor more parameterized models.

As mentioned, the model with the lowest information criteria is typically preferred; however, researchers have proposed cutoffs for noticeable improvements in model fit for certain information criteria. For example, Burnham and Anderson (2004) suggested that support for the model with a higher *AIC* is absent when the difference in *AIC* is greater than 10. Similarly, Kass and Raftery (1995) suggested that a *BIC* difference of 10 between two models provided very strong evidence favoring the model with a lower *BIC*; however, a *BIC* difference of less than two is negligible.

Alternatively, researchers have proposed descriptive quantifications of relative model fit for model selection (see Nagin, 1999; Masyn, 2013). For example, Kass and Wasserman (1995) have proposed using the Schwarz Information Criterion (*SIC*; $2 \cdot SIC = BIC$) for two competing models in order to calculate an *approximate Bayes Factor*, where the Bayes Factor is a ratio yielding the comparative likelihood that each model is the better fitting model. Additionally, there is the *approximate correct model probability*, which also uses the *SIC*, and the *AICc weight*, which normalizes the *AICc*, to estimate the probability that each of the fitted models is the best fitting model.

Approximate Likelihood Ratio Tests

The second group of model comparison statistics are approximate likelihood ratio tests. Mixture models with a different number of classes (e.g., *k* vs. $k-1$ classes), but the same set of parameter constraints are nested; however, the likelihood ratio test is not chi-squared distributed under the null hypothesis preventing its use. The likelihood ratio test is not chi-squared distributed because the parameter constraints applied to the *k* class model to create the $k-1$ class model are on the boundary of the parameter space. Thus, researchers have proposed modifications of the standard likelihood ratio test to statistically compare mixture models. The commonly used approximate likelihood ratio tests are based on the work of Vuong (1989) and Lo, Mendell, and Rubin (2001; LMR-LRT). The approximate likelihood ratio tests compare the fitted model with *k* classes to a similarly specified model with one fewer (i.e., $k-1$) class. The difference in the −2*LL* is computed and an associated *p*-value is estimated. If the *p*-value is less than the predetermined alpha level (i.e., 0.05), then the *k*-class model fits better than the $k-1$ class model. If the *p*-value is greater than alpha, then the fit of the two models is not considered to be statistically different, and the $k-1$ class model is preferred because the model is less parameterized.

The third approximate likelihood ratio test is the Bootstrap Likelihood Ratio Test (*BLRT*; McLachlan, 1987; McLachlan and Peel, 2000). The BLRT also compares the fitted model with *k*

classes to a similarly specified $k - 1$ class model. The difference in $-2LL$ is recorded and then data are simulated based on the parameter estimates from the $k - 1$ class model. The $k - 1$ and k class models are estimated using the simulated data to generate a sampling distribution for the difference in the $-2LL$ under the null hypothesis. The recorded difference in the $-2LL$ (from the empirical data) is then compared to the sampling distribution to estimate the p-value. As with the LMR-LRT, if the p-value is less than alpha (i.e., 0.05), then the k-class model is preferred to the $k - 1$ class model, and if the p-value is greater than alpha, then the fit of the two models is not statistically significant and the $k - 1$ class model is preferred.

Resampling Techniques

The third class of model fit statistics contains two approaches that are based on data resampling. First, Lubke and Campbell (2016) proposed a bootstrap approach to estimate model selection uncertainty in conjunction with information criteria (i.e., *AIC* and *BIC*). Here, the data are bootstrapped and the set of FMMs are fit to each bootstrapped sample. The information criteria are calculated and compared for each bootstrap sample. Given this information, researchers can evaluate both the sensitivity of each model's convergence and the sensitivity of model choice based on information criteria due to resampling. This information is particularly useful for model selection and provides uncertainty information that is sorely missing when comparing the fit of FMMs.

Second, Grimm et al. (2017) proposed a k-fold cross-validation approach to compare FMMs. In k-fold cross-validation (note the k in k-fold cross-validation is distinct from the k when referring to the number of classes [components] in FMMs), the data are randomly partitioned into k non-overlapping groups or folds. $k - 1$ folds are then used to estimate the model parameters for the FMMs. The kth fold is used for cross-validation and the FMMs estimated using $k - 1$ folds are applied to the kth fold and the $-2LL$ is retained. When the model is applied to the kth fold, parameters are not estimated, but fixed to the values obtained when the model was fit to the data from the $k - 1$ folds. Because the model is not estimated using the kth fold, the $-2LL$ does not have to be smaller for the more parameterized models (e.g., model with more classes). The k-fold cross-validation approach is performed k times with each fold serving as part of the estimation sample $k - 1$ times and serving as the validation sample one time. This provides a distribution of cross-validated $-2LL$s that can be used for model selection. The model with the lowest cross-validated $-2LL$ is selected or the simplest model with a cross-validated $-2LL$ within one standard error of the model with the lowest cross-validated $-2LL$ is selected. The k-fold cross-validation approach is the main approach to model selection in machine learning, where the performance of many models are compared.

As with Lubke's approach, the k-fold cross validation approach provides information on the sensitivity of the model's convergence because the model is estimated $k - 1$ times. If a model fails to converge for a portion of the $k - 1$ estimation attempts, then the model is no longer considered. Although it's important to note that sample size is slightly smaller when the model is estimated using k-fold cross-validation. Often k is set to 10 yielding 90% of the sample when the model is estimated (10% for the validation sample). If a greater portion of the sample is required to estimate each model, then k can be set to a higher value, such as 100 yielding 99% of the sample when each model is estimated. In the application of k-fold cross-validation for model selection with growth mixture models, a value of 10 and 100 for k yielded similar results and conclusions (Grimm et al., 2017).

Simulation Research on Model Comparison in Finite Mixture Models

Many simulation studies have been conducted to evaluate how the various model fit criteria for FMMs behave under a variety of population structures, statistical models (i.e., latent class, latent profile, growth mixture models), and sampling techniques (i.e., sample size, number of variables). The most often researched model fit criteria are information criteria, with fewer studies examining approximate likelihood ratio tests and resampling techniques. We review a sampling of this simulation work.

Fernández and Arnold (2016) recently examined model selection based on information criteria with ordinal data. Five versions of the *AIC* and two versions of the *BIC* were examined through simulation with sample sizes ranging from 50 to 500, with 5 or 10 variables, population structures with 2, 3, or 4 classes, and five different class configurations. The configurations attempted to create challenging scenarios for FMMs, where the classes overlapped. Fernández and Arnold (2016) found that the *AIC*, without adjustment, performed best, with an overall success rate of 93.8%. The *BIC* was less successful (43.7% accurate) and often underestimated the correct number of classes.

Six information criteria were examined by Yang (2006) through simulations with latent class models. The simulation study sampled data from 18 population structures with 4, 5, or 6 classes, and 12, 15, or 18 variables with sample sizes ranging from 100 to 1,000. Yang (2006) found that the *saBIC* performed best. The *AIC* and *BIC* both struggled with the *AIC* often overestimating the number of classes and the *BIC* underestimating the number of classes. Notably, the *BIC* performed reliably with a sample size of 1,000 and the *AIC* performed well with smaller sample sizes. Cubaynes et al. (2012) found a similar pattern of results to Yang (2006) regarding the *AIC* being liberal and the *BIC* being conservative when determining the number of classes in FMMs.

With regard to Yang's (2006) preference for the *saBIC*, Tofighi and Enders (2008) found support for the *saBIC* in their simulation work with growth mixture models (GMMs). In their simulation, Tofighi and Enders varied sample size (from $N = 400$–2000), number of repeated measures (4 and 7), class separation, relative class sizes, and the number of classes in the population. Tofighi and Enders (2008) found that the *saBIC* performed the best with a minimum success rate of 81%. Their findings for the *AIC* and *BIC* mimicked others as the *BIC*

performed poorly in small samples and the *AIC* performed poorly with larger samples. Tofighi and Enders also noted that the *BIC* performed poorly when class separation was low, even with a sample size of 1,000.

Yang and Yang (2007) drew similar conclusions in their simulation work with latent class models. Their simulations examined five versions of the *AIC* and four versions of the *BIC*, with sample sizes ranging from 200 to 1,000, with different population structures varying the number and relative size of the classes. Yang and Yang (2006) found that the *AIC* performed more poorly as sample size increased, and the *BIC* performed poorly when there was a large number of classes, particularly when sample size was small. Yang and Yang (2006) noted that the classes were not well separated in this condition where the *BIC* struggled; however, the *AIC* performed very well in the same condition.

Although, these researchers found mixed support for the *BIC*, other researchers have found broad support for the *BIC*. First, Nylund et al. (2007) performed extensive simulations of different types of FMMs, including latent class models, latent profile models, GMMs, and factor mixture models. In their latent class and latent profile analysis simulations, Nylund et al. (2007) varied sample size ($N = 200$, 500, and 1,000), the number of variables (8, 10, and 15), population structure (simple vs. complex), and the number of classes in the population (3 and 4). They found that the *AIC* rarely found the correct number of classes in any setting and often overestimated the number of classes, whereas the *BIC* performed well across a variety of conditions, but struggled with the ten item LCA with $N = 200$. We note that the classes were well separated in Nylund et al. (2007), which likely affected their findings. Second, Steele and Raftery's (2009) simulation research on univariate FMMs found broad support for the *BIC*. In their simulations, sample size ranged from 100 to 400 with one and two class models. The *BIC* successfully identified the correct number of classes in 94% of the simulations, outperforming the *AIC* and other model fit criteria. However, the *BIC* struggled when there were two classes and the classes were not well separated. The *AIC*, as with Yang (2006) and Cubaynes et al. (2012) work, frequently overestimated the true number of classes.

Simulation work focused on GMMs have found mixed results with respect to the performance of information criteria. Grimm et al. (2013) found that information criteria generally performed poorly – accurately determining a two-class population structure in less than 20% of the replicates. However, important associations were found that between the simulation conditions and the likelihood of the information criteria favoring the two-class model. The *BIC* was most sensitive to class separation. For example, the *BIC* was ten times more likely to favor the two-class model when the mean difference in the intercept or slope of the two classes was three standard deviations apart compared to when they were two standard deviations apart. The *BIC* was also sensitive to sample size, the relative size of the two classes, the number of repeated measurements, and the location of the class differences (intercept vs. slope). Similarly, Peugh and Fan (2013) found that information criteria performed poorly across a variety of circumstances, and observed a strong association between the performance of the information criteria and sample size. Peugh and Fan (2013) attributed the poor performance of information criteria to two factors – sample size and residual variability. Peugh and Fan (2013) noted that Paxton et al. (2001) suggested that a sample size of less than 500 was inadequate for heterogenous structural equation models. The second factor, residual variability, has the ability to mask class differences even though the amount of residual variability was not excessive in their simulations.

Compared to the amount of research into information criteria for model selection in FMMs, there are few studies that have examined approximate likelihood ratio tests. Nylund et al. (2007) found that *BLRT* outperformed information criteria across a range of models, Tofighi and Enders (2008) found adequate performance of the *LMR–LRT*, and Grimm et al. (2013) found that the *LMR–LRT* outperformed information criteria across a range of simulation conditions. Grimm et al. (2013) highlighted how the *LMR–LRT* was sensitive to several simulation conditions including sample size, class separation, relative class sizes, and the location of the differences (intercept vs. slope).

Simulation research on resampling techniques is even more limited. He and Fan (2019) evaluated the proposed approaches using *k*-fold cross-validation for model selection and found that these approaches struggled to identify the proper number of classes. Finally, Lubke et al. (2017) performed simulation research to examine the benefit of using bootstrap samples when evaluating class enumeration. While the bootstrap samples did not lead to a model selection criterion, Lubke et al. (2017) found that the bootstrap samples can aid model selection compared to using the *AIC* and *BIC* alone.

Benefits and Limitations of Model Fit Approaches

The conclusions from the simulation research and the varied recommendations for model comparison with FMMs suggest that the conclusions were strongly dependent on the simulation conditions considered (Grimm et al. 2017). For example, Fernández and Arnold (2016) focused their simulations on mixture components that were not very distinct, and their results favored the *AIC* – an information criterion that minimally penalizes parameters. Because of the different recommendations, it's important to consider the benefits and limitations of the different model fit criteria.

Information criteria are attempting to appropriately balance the information available from how well the model captures the data in terms of the $-2LL$ and the penalty for the number of estimated parameters. The size of the $-2LL$ is dependent on 1) model fit – the match (or mismatch) between the model's expectations and the data, 2) the sample size, and 3) the number of variables. Given constant model fit, the $-2LL$ changes in a linear fashion with sample size and the number of variables. **Figure 1** contains two plots highlighting these associations with simulated data from a latent profile model. The data were generated for two classes with a one standard deviation difference in the mean of each variable between classes. Given the population structure, the correlations between

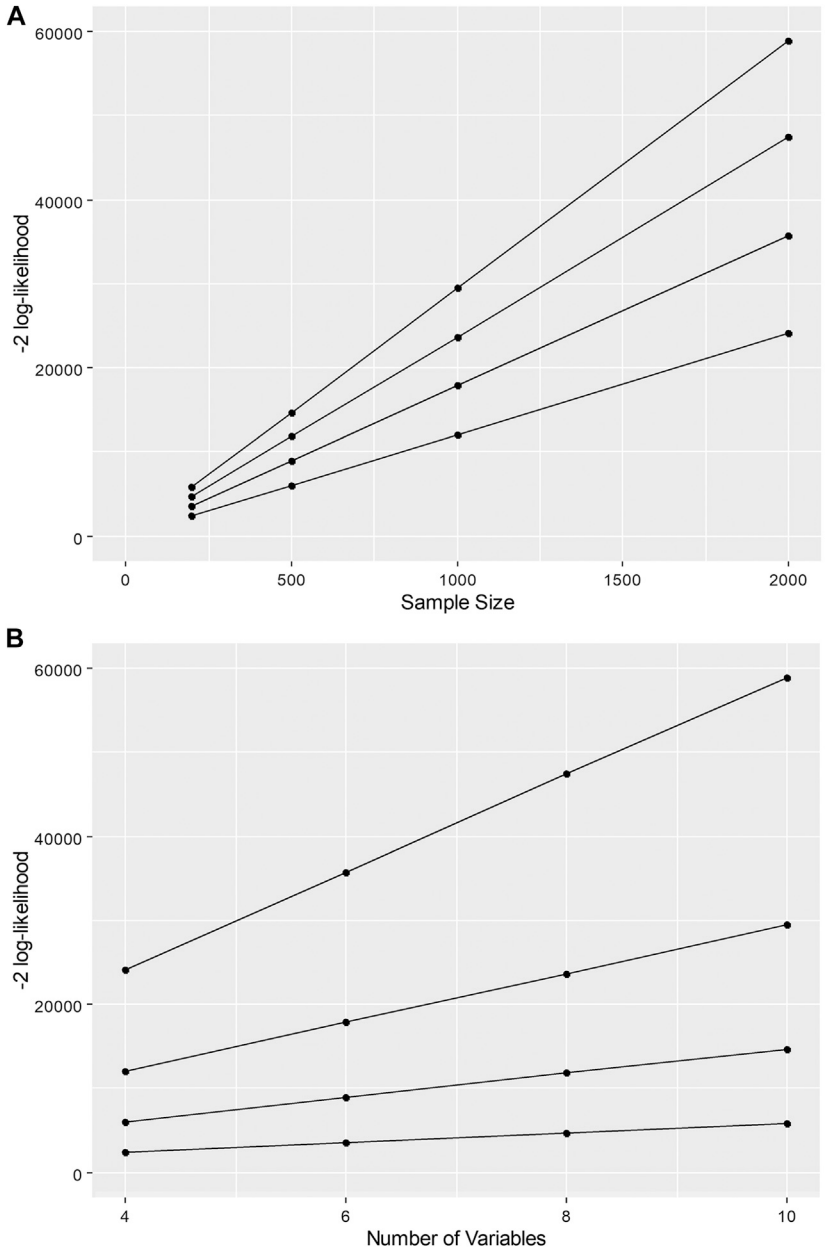

FIGURE 1 | −2 log-likelihood as a function of **(A)** sample size and **(B)** number of variables in latent profile analysis given a constant difference in a two-class model.

variables conditioned on class membership are near zero, which can affect these simple conclusions. One hundred samples were drawn from the population, a 2-class latent profile model was estimated, and the average $-2LL$ is plotted. The $-2LL$ is directly proportional to the sample size: the averaged $-2LL$ with $N = 2,000$ is approximately twice the averaged $-2LL$ with $N = 1,000$. The effect for sample size is slightly smaller: the average $-2LL$ for the eight variable models is 1.97 times larger than the average $-2LL$ from the four variable models.

These associations highlight the challenge for information criteria when comparing mixture models. For example, given a constant difference in model fit between two competing models, the difference in the $-2LL$ is linearly related to sample size; however, the penalty for each parameter is constant in the *AIC* and nonlinearly associated with sample size in the *BIC* (and variants of the *AIC* and *BIC*). Thus, the simulation studies that keep the number of variables and population model consistent, but vary sample sizes invariably determine that the *AIC* performs better than the *BIC* in smaller samples and the *BIC* performs better than the *AIC* in larger samples. Similar findings are present when considering the number of variables (see Nylund et al. 2007) and model discrepancy (see Grimm et al., 2013), where the *AIC* performed better when the classes were not very distinct (but overestimates the number of classes when the classes are more

distinct) and the *BIC* performed better when the classes were more distinct.

A second challenge for information criteria is missing data. Specifically, for many information criteria (e.g., *BIC*, *saBIC*, *AICc*), the penalty for each parameter is dependent on sample size, which is a marker for the amount of information available in the data. However, when missing data are present, there is less information in the data for the same sample size (assuming the number of variables is constant). For example, in our simulated 2-class latent profile data (**Figure 1**), the $-2LL$ changed linearly with respect to the proportion of missing data, such that the average $-2LL$ for the full sample was 1.97 times larger than the average $-2LL$ when 50% of the data were randomly deleted (the difference in $-2LL$ can be nonlinear depending on the model and population structure). Thus, missing data have a similar effect as sample size on the $-2LL$ and therefore, the *AIC* is likely to perform better when more missing data are present and the *BIC* is likely to perform better when data are more complete. Given these challenges, it is clear that there will not be a single per parameter penalty that will lead to proper model selection across a wide range of FMMs, sample sizes, number of variables, and missing data.

Approximate likelihood ratio tests have shown tremendous promise in simulation studies and provide a statistical framework for comparing FMMs. We identify two limitations of approximate likelihood ratio tests. The first is that the approximate likelihood ratio tests are only available when comparing FMMs that differ by one class with the same set of model constraints. Thus, the approximate likelihood ratio tests are not available when comparing a 2-class model with a 4-class model, two 2-class models with different constraints, or a 2 and 3-class model with different constraints. The second limitation for the approximate likelihood ratio tests is a limitation shared by all statistical tests, which is the association between the value of the likelihood ratio test and sample size. When working with large sample sizes, the approximate likelihood ratio tests can be significant suggesting the model with more classes fits significantly better than the model with fewer classes even when the difference in model fit is small and meaningless.

The resampling techniques provide important information about the sensitivity of estimating the model's parameters and the replicability of the model through repeated estimation using bootstrap samples or *k*-fold cross-validation. This information is not contained in other model fit statistics and is important because FMMs often experience convergence issues. Compared to the other model fit statistics, simulation research on these resampling techniques for FMMs is limited and more research is greatly needed. He and Fan (2019) indicated challenges for the *k*-fold cross-validation approaches; however, He and Fan (2019) noted that *k*-fold cross-validation was not strongly related to sample size – a potential plus compared to other model fit criteria.

Given the challenges and limitations of the available model fit criteria, we take an alternative point of view when comparing FMMs. The approach we consider looks at model comparison through the lens of *effect sizes* to provide context about differences in model fit. In the next section we outline the effect size measure and discuss how it can aid the evaluation of model comparison in FMMs.

AN EFFECT SIZE MEASURE FOR MODEL COMPARISON

Effect sizes quantify the magnitude of a parameter or the difference in two entities. For example, Cohen's **d** (Cohen, 1988) is an effect size measure when comparing means for two groups of individuals. Cohen's **d** is defined as

$$\mathbf{d} = \frac{\overline{y}_1 - \overline{y}_2}{s} \tag{6}$$

where $\overline{y}_1$ and $\overline{y}_2$ are the estimated means of the outcome for groups 1 and 2, respectively, and s is the pooled standard deviation. Cohen's **d** is the standardized mean difference and provides information about the magnitude of the difference between groups and is relatively unaffected by sample size (Cohen's **d** is generally unrelated to sample size, but can be biased in small samples). While Cohen's **d** does not replace statistical tests for group differences (e.g., *t*-test), it adds to the discussion about the meaning of statistically significant group differences. For example, an intervention leads to a statistically significant effect on participant depression, but if the effect size is quite small, then further investment into the intervention may not be warranted.

In multiple regression analysis with numeric predictors, the commonly reported effect size is the *standardized beta weight*. In linear regression, the standardized beta weight is calculated as

$$\beta_1 = b_1\left(\frac{s_{x1}}{s_y}\right) \tag{7}$$

where β_1 is the standardized beta weight for the predictor variable x_{1i}, b_1 is the estimated unstandardized regression coefficient for x_{1i}, s_{x1} is the estimated standard deviation of x_{1i}, and s_y is the estimated standard deviation of the outcome y_i. Again, the standardized beta weight is a rescaling of the unstandardized regression coefficient to put its magnitude into context. Here, the scaling is in terms of the ratio of the standard deviations of the predictor and outcome.

For determining an effect size measure for model comparison, our goal is to scale the difference in the $-2LL$ relative to a standard. One way to do this is by dividing the difference in $-2LL$ by the $-2LL$ of the less parameterized model. This fit statistic was first proposed by McArdle et al. (2002) and termed the *likelihood increment percentage* (*LIP*) and calculated as

$$LIP_{model(j)} = 100 \cdot \left(1 - \frac{-2LL_{model(j)}}{-2LL_{baseline}}\right) \tag{8}$$

where $-2LL_{model(j)}$ is the $-2LL$ for the more parameterized model and $-2LL_{baseline}$ is the $-2LL$ for the less parameterized (baseline) model. The *LIP* yields the percent improvement in the $-2LL$ from the baseline model for the more parameterized model. The same metric, but scaled in terms of the proportional improvement in

model fit, has been implemented in recursive partitioning algorithms for mixed-effects (Abdolell et al., 2002; Stegmann et al., 2018) and structural equation models (Serang et al., 2020).

The main reason to do this is to provide context regarding the relative improvement in model fit. The examination of model comparison indices, such as information criteria, through simulation is focused on recovering the number of classes that are known to exist within the data. Even though researchers control the distinctiveness of the classes, we don't have an appropriate metric for this distinctiveness. When fit indices fail to uncover the proper number of classes, the blame is put on the fit index. If the same approach were taken in simulation studies for the analysis of variance, we may lead to the conclusion that null hypothesis significance testing (NHST) isn't working. For example, NHST is unable to consistently find a significant difference between two means when Cohen's **d** = 0.2 in the population with $N = 200$. Instead of blaming NHST, we conclude that we lack statistical power because Cohen's **d** = 0.2 is a small effect size. In FMMs and multivariate models more generally, we don't have a standard effect size metric for model comparison. Thus, one of our goals is to define small, medium, and large effect sizes for FMMs based on the *LIP*.

The second reason for considering the *LIP* for model comparison is because it should be weakly related to sample size because it is scaled by the $-2LL$ of the less parameterized model. We identified sample size as a challenge for information criteria and approximate likelihood ratio tests because of its association with the $-2LL$ and the change in $-2LL$ given the same magnitude of differences between the classes.

To determine how the *LIP* changes as a function of different latent class structures and sample sizes, we mimic two previously conducted simulation studies. The first is Nylund et al.'s (2007) simulation study for latent profile analysis and the second is Grimm et al.'s (2013) simulation study for growth mixture models.

LIP EFFECT SIZES

Latent Profile Models

Simulation research following the population models for latent profile analysis (latent class analysis with continuous indicators) in Nylund et al. (2007) was conducted to examine the *LIP*. For the *LIP*, we examined the percent improvement in the $-2LL$ for adding one class. That is, when comparing the 2 and 3-class models, the 2-class model served as the baseline model and the 3-class model served as the model being evaluated, and when comparing the 3 and 4-class models, the 3-class model served as the baseline model and the 4-class model served as the model being evaluated. We also report model selection based on the *AIC*, *BIC*, and *saBIC*.

Nylund et al. (2007) examined three latent profile population structures (8-variable simple, 15-variable simple, and 10-variable complex) under three sample sizes ($N = 200$, 500, and 1,000). In **Table 1** we report the mean and standard deviation of the *LIP* across replicates. We also report model choice based on the *AIC*, *BIC*, and *saBIC* as the proportion of replicates in which each model had the lowest information criteria. Following Nylund et al. (2007), the *BIC* performed very well, and the *AIC* and *saBIC* performed admirably, but tended to overestimate the number of classes. Interestingly, the *AIC* tended to overestimate the number of classes as sample size increases, whereas the *saBIC* tended to overestimate the number of classes in the smaller sample size conditions.

The mean *LIP* and the standard deviation of *LIP* values across replicates are contained on the right-hand side of **Table 1**. The first LPA structure was the 8-variable simple structure population model, which had four classes. The mean *LIP* comparing the 1 to 2, 2 to 3, and 3 to 4-class models were all over 2; however, the mean *LIP* comparing the 4- to 5-class model was 0.30 or smaller. For all sample sizes in the 8-variable population structure, there was a clear drop in the *LIP* after the 4-class model. The *LIP* values were fairly consistent across sample sizes, as was expected for effect size type measures. In this 8-variable population structure, there were two items for each class with a mean of 2 and all remaining items had means of 0 (with standard deviations of 1) to indicate the distinctiveness of the classes in this simulation condition.

The second population structure was the 15-variable simple structure, where there were three classes in the population. Each class had five variables with a mean of 2 with remaining variables having a mean of 0 (with a standard deviation of 1). Again, there were clear changes in the *LIP* after the 3-class model. *LIP* values were greater than six when comparing the 1 to 2-class model and when comparing the 2 to 3-class model. The average *LIP* was less than 0.30 when comparing the 3 and the 4-class model. Thus, the *LIP* clearly delineated the population structure. Again, the classes were well separated in this population structure. The higher-values of the *LIP*, compared to the 8-variable population structure, reflect two things. First, the classes in the 15-variable simple population structure were more distinct. That is, five variables distinguished each class compared to two-variables in the 8-variable simple population structure. Second, the difference in the number of estimated parameters when increasing the number of classes. Specifically, nine parameters were added when increasing the number of classes by one in the 8-variable simple population structure, and 16 parameters were added when increasing the number of classes in the 15-variable simple population structure. This second reason suggests determining the percent improvement in model fit *per additional parameter*. Dividing *LIP* values by the difference in the number of additional parameters makes them more comparable. For example, dividing the 4-class *LIP* values for the 8-variable simple structure by nine yields 0.37, and dividing the 3-class *LIP* values for the 15-variable simple structure by 16 yields 0.53.

The third and final LPA population structure in Nylund et al. (2007) was the 10-variable complex structure. In this population structure there were four classes, and no one variable distinguished each class. There was a class with a mean of two for all variables, a class with a mean of two for the first five variables and a mean of zero for the second five variables, a class with a mean of zero for the first five variables

TABLE 1 | Model selection in latent profile structures from Nylund et al. (2007) and associated *LIP* values.

	AIC					*BIC*					saBIC					*LIP*				
N	**2**	**3**	**4**	**5**	**6**	**2**	**3**	**4**	**5**	**6**	**2**	**3**	**4**	**5**	**6**	**2**	**3**	**4**	**5**	**6**
8-Variable simple																				
200	0	0	**69**	23	8	0	0	**100**	0	0	0	0	**78**	18	4	2.27 (0.42)	2.63 (0.45)	**3.27 (0.50)**	0.30 (0.10)	0.26 (0.11)
500	0	0	**73**	20	7	0	0	**100**	0	0	0	0	**99**	1	0	2.12 (0.24)	2.44 (0.25)	**3.27 (0.29)**	0.12 (0.04)	0.11 (0.04)
1,000	0	0	**72**	17	11	0	0	**100**	0	0	0	0	**100**	0	0	2.04 (0.17)	2.37 (0.19)	**3.30 (0.22)**	0.06 (0.02)	0.06 (0.02)
15-Variable simple																				
200	0	**85**	13	2	0	0	**100**	0	0	0	0	**91**	8	1	0	6.48 (0.42)	**8.55 (0.55)**	0.27 (0.07)	0.26 (0.07)	0.21 (0.09)
500	0	**83**	14	3	0	0	**100**	0	0	0	0	**100**	0	0	0	6.29 (0.29)	**8.45 (0.33)**	0.12 (0.03)	0.11 (0.03)	0.08 (0.04)
1,000	0	**74**	20	4	2	0	**100**	0	0	0	0	**100**	0	0	0	6.22 (0.21)	**8.45 (0.23)**	0.06 (0.01)	0.06 (0.01)	0.05 (0.01)
10-Variable complex																				
200	0	1	**62**	14	23	0	16	**84**	0	0	0	1	**67**	12	20	4.88 (0.64)	3.05 (0.67)	**1.40 (0.52)**	0.31 (0.10)	0.23 (0.37)
500	0	0	**74**	20	6	0	0	**100**	0	0	0	0	**99**	1	0	4.82 (0.41)	2.86 (0.41)	**1.32 (0.28)**	0.12 (0.04)	0.11 (0.04)
1,000	0	0	**44**	17	39	0	0	**100**	0	0	0	0	**92**	1	7	4.79 (0.29)	2.78 (0.30)	**1.31 (0.18)**	0.07 (0.02)	0.06 (0.10)

Columns represent latent profile models with a different number of classes, values for the information criteria represent the percent of replicates where the model had the lowest information criteria, AIC = *Akaike Information Criterion,* BIC = *Bayesian Information Criterion,* saBIC = *Sample Size Adjusted Bayesian Information Criterion,* LIP = *Likelihood Increment Percentage,* LIP *values are mean values across replicates with the standard deviation across replicates within parentheses for comparing the model with one fewer class. Bolded columns indicate the number of classes in the population.*

TABLE 2 | Model Selection in Growth Mixture Modeling Structures with intercept differences, five time points, and 50–50 mixing proportion from Grimm et al. (2013) and Associated *LIP* values.

	AIC			*BIC*			*saBIC*			*LIP*	
N	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**2**	**3**
1. Standard deviation difference in intercepts											
200	65	**23**	12	100	**0**	0	73	**18**	9	**0.08 (0.05)**	0.07[a] (0.06)
500	71	**17**	11	100	**0**	0	92	**8**	0	**0.03 (0.02)**	0.03 (0.02)
1,000	76	**15**	9	100	**0**	0	96	**3**	1	**0.01 (0.01)**	0.01 (0.01)
2. Standard deviation difference in intercepts											
200	50	**40**	10	95	**5**	0	54	**37**	**9**	**0.12 (0.08)**	0.08 (0.05)
500	19	**66**	16	91	**9**	0	47	**49**	**4**	**0.07 (0.04)**	0.02 (0.02)
1,000	2	**80**	18	70	**30**	0	18	**80**	**3**	**0.06 (0.02)**	0.01 (0.01)
3. Standard deviation difference in intercepts											
200	0	**80**	20	26	**73**	1	1	**81**	18	**0.38 (0.15)**	0.07 (0.05)
500	0	**82**	18	0	**100**	0	0	**94**	6	**0.34 (0.09)**	0.02 (0.02)
1,000	0	**86**	14	0	**100**	0	0	**98**	2	**0.34 (0.06)**	0.01 (0.01)

Columns represent latent profile models with a different number of classes, values for the information criteria represent the percent of replicates where the model had the lowest information criteria, AIC = *Akaike Information Criterion,* BIC = *Bayesian Information Criterion,* saBIC = *Sample Size Adjusted Bayesian Information Criterion,* LIP = *Likelihood Increment Percentage,* [a]*based on 99% of the replicates because model converged at a local maxima in the likelihood function,* LIP *values are mean values across replicates with the standard deviation across replicates within parentheses for comparing the model with one fewer class. Bolded columns indicate the number of classes in the population.*

and a mean of two for the second five variables, and a class with a mean of 0 for all ten variables. Additionally, class sizes were unequal with relative class sizes being 5, 10, 15, and 70% (note Nylund et al. (2007) reports 75% in the fourth class). As with the other two population structures, the mean *LIP* clearly indicates when more classes were not warranted. The mean *LIP* was 4.8 when comparing the 2 and 1-class models, 2.9 when comparing the 3 and 2-class models, and 1.4 when comparing the 4 and 3-class models. The comparison of the 5-class model to the 4-class model yielded a mean *LIP* that was less than 0.4 – a similar value was obtained in the other population structures. The class differences were not as distinct in this third population structure, which is highlighted by the smaller *LIP* values when moving from the 3-class model to the 4-class model. Dividing the *LIP* by the difference in the number of estimated parameters yields 0.12. Noticeably smaller than the other two population structures where the class differences were much larger.

Growth Mixture Models

The growth mixture modeling simulations by Grimm et al. (2013) focused on two-class mixture models and varied sample size (N = 200, 500, 1,000), the number of measurement occasions (T = 5, 7, 9), the mixing proportion (50–50, 80–20, 95–5), whether the class differences were in the intercept or the slope, and the magnitude of the differences between the classes (1, 2, 3 standard deviations). The simulations we replicated focused on intercept differences, a mixing proportion of 50–50, and varied sample size; however, we also discuss simulations involving the number of measurement

TABLE 3 | Model selection in growth mixture modeling structures with a three standard deviation difference in the intercepts from Grimm et al. (2013) and Associated *LIP* values.

	AIC			*BIC*			*saBIC*			*LIP*	
N	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**2**	**3**
Timepoints = 5 and 80-20 mixing proportion											
200	1	**77**	22	18	**82**	0	1	**79**	20	**0.43 (0.17)**	0.07 (0.05)
500	0	**76**	24	1	**99**	0	0	**89**	11	**0.40 (0.10)**	0.03 (0.02)
1,000	0	**82**	18	0	**100**	0	0	**99**	1	**0.40 (0.08)**	0.01 (0.01)
Timepoints = 5 and 95-5 mixing proportion											
200	11	**65**	24	63	**37**	0	14	**65**	**21**	**0.25 (0.16)**	0.08 (0.06)
500	1	**72**	27	14	**86**	0	3	**90**	**7**	**0.22 (0.08)**	0.03 (0.02)
1,000	0	**78**	22	1	**99**	0	0	**97**	**3**	**0.21 (0.06)**	0.01 (0.01)
Timepoints = 7 and 80-20 mixing proportion											
200	4	**78**	18	93	**7**	0	6	**80**	15	**0.41 (0.12)**	0.05 (0.03)
500	0	**78**	22	19	**81**	0	0	**92**	8	**0.34 (0.07)**	0.02 (0.01)
1,000	0	**84**	16	0	**100**	0	0	**97**	3	**0.32 (0.05)**	0.01 (0.01)
Timepoints = 9 and 80-20 mixing proportion											
200	0	**76**	24	9	**91**	0	0	**80**	20	**0.27 (0.10)**	0.04 (0.03)
500	0	**81**	19	0	**100**	0	0	**98**	2	**0.26 (0.06)**	0.01 (0.01)
1,000	0	**82**	18	0	**100**	0	0	**99**	1	**0.25 (0.04)**	0.01 (0.01)

Columns represent latent profile models with a different number of classes, values for the information criteria represent the percent of replicates where the model had the lowest information criteria, AIC = *Akaike Information Criterion,* BIC = *Bayesian Information Criterion,* saBIC = *Sample Size Adjusted Bayesian Information Criterion,* LIP = *Likelihood Increment Percentage,* LIP *values are mean values across replicates with the standard deviation across replicates within parentheses for comparing the model with one fewer class. Bolded columns indicate the number of classes in the population.*

occasions and the mixing proportion with classes separated by three standard deviations. The 1, 2, and 3-class models were fit with the means of the intercept and slope allowed to vary over the classes. Thus, three additional parameters were estimated when increasing the number of classes by one. Of the models specified, the 2-class model is consistent with the population structure.

Table 2 contains model selection percentages based on the *AIC*, *BIC*, and *saBIC*, as well as the mean and standard deviation of the *LIP* values when comparing the 1 and 2-class models and when comparing the 2 and 3-class models. All of the information criteria struggled to properly identify the 2-class model when the intercept means for the two classes were one standard deviation apart (top portion of **Table 2**). The *LIP* also struggled because *LIP* values were small and were unable to differentiate the comparison of the 2-class model to the 1-class model and the 3-class model to the 2-class model. These findings are consistent with Rindskopf's (2003) demonstration that even when two classes are different by one standard deviation, the data appear normal suggesting the data came from a single population.

In the middle section of **Table 2**, the intercept means were two standard deviations apart. Here, the *AIC* and *saBIC* performed well with $N = 1,000$; however, these indices performed poorly in the smaller sample size conditions. The *BIC* performed poorly across all sample sizes with this degree of class separation. The mean *LIP* showed a clear advantage for the 2-class model with $N = 500$ and $N = 1,000$, but struggled with $N = 200$. When the intercept means were three standard deviations apart (bottom of **Table 2**), the information criteria performed well with the *BIC* leading the way and the mean *LIP* values clearly indicated a preference for the 2-class model.

Table 3 contains four more population structures. In all of these population structures, the means of the intercepts were three standard deviations apart. In the first two, the mixing proportion was 80–20 and 95–5, respectively. In Grimm et al. (2013), the 80–20 mixing proportion led to slightly better model selection rates for the information criteria, and the 95–5 mixing proportion led to slightly worse model selection rates compared to when the mixing proportion was 50–50. These patterns held for these simulations. Compared to when there was a 50–50 mixing proportion, the *LIP* values were slightly greater for the 80–20 mixing proportion (~0.40) and slightly lower for the 95–5 mixing proportion (~0.21), which was consistent with the performance of the information criteria.

In the second two population structures, the number of measurement occasions was changed to seven and nine, respectively. In these population structures, the intercept means were three standard deviations and the mixing proportions were 80–20. The effect of the number of time points in this population structure was unexpected. The information criteria correctly identified the two-class models slightly more with nine measurement occasions, compared to the five time point data; however, the *BIC* struggled when there were seven measurement occasions with $N = 200$. Interestingly, the *LIP* indicated that the effect size was smaller when there were more time points; however, this finding can be explained. The difference between the two classes was in the intercept means. Given the non-zero variance in the slope for each class, the observed score variance increased over time. Thus, the standardized difference in the means between the two classes was smaller at the later time points, which is why the *LIP* decreased with more measurement occasions with this population structure. To further examine the effect of the number of time points on the 2-class models, we examined the mean entropy. Entropy was lowest for the 2-class model with $T = 5$ at 0.779, highest for the 2-class model with $T = 7$ at

0.837, and 0.801 for the 2-class model with $T = 9$. Interestingly, entropy was highest when there were seven time points, indicating that classification quality was highest for this population structure.

Interim Summary

Overall, the *LIP* showed a clear drop when no more classes were warranted for all of the latent profile models. The drop in the *LIP* was less clear for the growth mixture models. In the growth mixture models, the distinctiveness of the two classes was varied systematically. The *LIP* did not show a systematic change after the two-class model when the classes differed by one standard deviation in the intercept mean. There was a small drop when the intercept means were two standard deviations apart, and a clear drop when the intercept means were three standard deviations apart. In the growth mixture modeling simulations, the information criteria performed poorly, which highlights the challenges in comparing FMMs.

Although the performance of the *LIP* was admirable, there were two concerns. First, the *LIP* was mildly associated with sample size. That is, the *LIP* tended to be higher with $N = 200$ compared to $N = 500$ and $N = 1,000$. Thus, the *LIP* is likely to be overestimated with smaller samples. Second, the *LIP* varied across replicate simulations. While this was expected, the variance was fairly large, particularly in small samples. We therefore caution its use with $N < 500$.

LIP Effect Sizes

The performance of the *LIP* leads to the question of *how* the *LIP* should be used for model comparison. We propose first examining how the *LIP* changes when increasing the number of classes. Ideally, the *LIP* will drop off when no more classes are warranted – akin to scree plot in factor analysis. Second, we recommend dividing the *LIP* by the increase in the number of estimated parameters when a class is added to the model. This *LIP per parameter* (*LIPpp*) takes model complexity into account.

The *LIPpp* was approximately 0.37 in Nylund et al. (2007) 8-variable simple LPA. We consider this to be a *large* effect size for the *LIPpp*. For this effect size, the *BIC* performed well across all sample sizes considered. Given the variability in *LIP* across replicates, we consider *LIPpp* values greater than 0.30 to indicate a large improvement in model fit. The *LIPpp* was approximately 0.12 in Nylund et al. (2007) 10-variable complex LPA when comparing the 3-class model to the 2-class model, and 0.11 in Grimm et al. (2013) simulations when the intercept means were three standard deviations apart. We consider these to be *medium* effect sizes for the *LIPpp*. For this effect size, the *BIC* performed well when sample sizes were 500 and 1,000. Given the variability in *LIP* across replicates, we consider *LIPpp* values between 0.10 and 0.30 to indicate a medium improvement in model fit. Finally, the *LIPpp* was 0.02 in Grimm et al. (2013) simulations when the intercept means were two standard deviations apart. We consider this to be a *small* effect size. Here, the *AIC* performed well with a large sample size (i.e., $N = 1,000$), but the *BIC* struggled given the sample sizes considered. We therefore consider *LIPpp* values between 0.02 and 0.10 to represent small improvements in model fit. We recognize that these effect sizes for the *LIPpp* are raw and require further study, and are solely proposed for FMMs.

DISCUSSION

Model comparison in FMM is challenging and existing model fit indices, such as information criteria, approximate likelihood ratio tests, and those based on resampling techniques do not provide standardized information about the relative improvement in model fit. The *LIP*, initially proposed by McArdle et al. (2002) for model comparison, was divided by the difference in the number of estimated parameters and is a proposed *effect size* to aid model comparison for FMMs. This *LIPpp* provides a measure of the relative improvement in model fit per parameter when comparing two models. Different values of the *LIPpp* were associated with small to large differences in model fit, and these values may help to provide context when comparing two models as opposed to solely examining the statistical significance of the difference in model fit, or examining which model had lower information criteria. We found that the *LIPpp* was slightly inflated with $N = 200$ and caution against its use with smaller sample sizes.

When comparing two models, researchers have categorized differences in the *AIC* (Burnham and Anderson, 2004) and *BIC* (Kass and Raftery, 1995), and proposed an approximate Bayes Factor, to provide context regarding the magnitude of the difference in model fit. While not the same as an effect size, the contextual comparison is helpful when comparing two (or more) models. The *LIPpp* provides more direct information about the magnitude of improvement in model fit because it is less influenced by sample size than the *AIC* and *BIC*, and takes the difference in the number of estimated parameters into account.

Mahalanobis Distance and Entropy

There are two important pieces of statistical information from FMMs that are relevant when discussing the *LIPpp* – *Mahalanobis distance* (Mahalanobis, 1936) and *entropy*. Mahalanobis distance is a measure of the distance between two mean vectors standardized by a common or pooled covariance matrix. Mahalanobis distance has been utilized as an *effect size* measure for the separation of two classes in FMM simulations (Grimm, et al., 2013; Peugh and Fan, 2012, Peugh and Fan, 2015). The multivariate Mahalanobis distance is calculated as

$$D = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \quad (9)$$

where μ_1 and μ_2 mean vectors for the first and second classes, respectively, and Σ is the pooled expected covariance matrix within each class. Mahalanobis distance indicates how distinct two classes are from one another. Mahalanobis distance is an effect size measure; however, it is distinct from the *LIPpp*. When comparing a two-class model to a one-class model, Mahalanobis distance and the *LIPpp* will be highly related for a given model (e.g., growth mixture model). However, Mahalanobis distance is more limited when moving to models with more classes because it

is used to make pairwise comparisons, and is independent of the difference in the number of estimated parameters across models.

Entropy is a measure of classification quality and is calculated as

$$entropy = 1 + \frac{1}{N\log(\mathrm{K})}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{ik}\left(\log\left(p_{ik}\right)\right) \quad (10)$$

where N is the sample size, K is the number of classes, and p_{ik} is the posterior probability that person i is a member of class k. While we expect entropy to be associated with *LIPpp*, the two statistics provide different information. Entropy is not a model fit statistic, but a characteristic of a fitted FMM. Thus, entropy and *LIPpp* serve different purposes. We expect entropy to be weakly related to *LIPpp*. That is, a model with k classes may have high entropy and fit much better than the model with $k-1$ classes (high *LIPpp*); however, it is also the case that a model with k classes may have high entropy and not fit much better than the model with $k-1$ classes (low *LIPpp*).

Concluding Remarks

The *LIPpp* was proposed as a measure of relative improvement in model fit for comparing FMMs. In our simulation work, the *LIPpp* was able to discern the number of classes relatively well. The *LIPpp* plateaued when increasing the number of classes was not warranted. In our simulations, we focused on models with the same constraints, but a different in the number of classes; however, the *LIPpp* could be applied when comparing mixture models that differ in the number of classes and/or class constraints. Future research should evaluate the *LIPpp* in other contexts where a sequence of models is compared, such as when studying factorial invariance. Additionally, given the variability in the *LIPpp* across replicate samples in our simulation, the *LIPpp* should be examined in the context of resampling techniques (e.g., bootstrapping) to obtain a distribution of *LIPpp* values when increasing the number of latent classes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

KG developed the idea, conducted simulation work, and drafted the paper. RH conducted the literature, helped draft the introduction, and edited the paper. DR helped draft the introduction and edited the paper.

REFERENCES

Abdolell, M., LeBlanc, M., Stephens, D., and Harrison, R. V. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat. Med.* 21, 3395–3409. doi:10.1002/sim.1266

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *2nd International symposium on information theory*. Editors B. N. Petrov and F. Csáki (Budapest, Hungary: Akadémiai Kiadó), 267–281.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33, 261–304. doi:10.1177/0049124104268644

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers, 567.

Cubaynes, S., Lavergne, C., Marboutin, E., and Gimenez, O. (2012). Assessing individual heterogeneity using model selection criteria: how many mixture components in capture-recapture models?: heterogeneity, mixtures and model selection. *Methods Ecol. Evol.* 3, 564–573. doi:10.1111/j.2041-210X.2011.00175.x

Fernández, D., and Arnold, R. (2016). Model selection for mixture-based clustering for ordinal data. *Aust. N. Z. J. Stat.* 58, 437–472. doi:10.1111/anzs.12179

Gibson, W. A. (1959). Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika* 24, 229–252. doi:10.1007/BF02289845

Grimm, K. J., Mazza, G. L., and Davoudzadeh, P. (2017). Model selection in finite mixture models: a *k*-fold cross-validation approach. *Struct. Equ. Model.* 24, 246–256. doi:10.1080/10705511.2016.1250638

Grimm, K. J., Ram, N., Shiyko, M. P., and Lo, L. L. (2013). "A simulation study of the ability of growth mixture models to uncover growth heterogeneity," in *Contemporary issues in exploratory data mining*. Editors J. J. McArdle and G. Ritschard (New York, NY: Routledge Press), 172–189.

He, J., and Fan, X. (2019). Evaluating the performance of the *k*-fold cross-validation approach for model selection in growth mixture modeling. *Struct. Equ. Model.* 26, 66–79. doi:10.1080/10705511.2018.1500140

Hurvich, C. M., and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297–307. doi:10.1093/biomet/76.2.297

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi:10.1080/01621459.1995.10476572

Kass, R. E., and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* 90, 928–934. doi:10.1080/01621459.1995.10476592

Lazarsfeld, P. F. (1950). "The logical and mathematical foundation of latent structure analysis and the interpretation and mathematical foundation of latent structure analysis," in *Measurement and prediction*. Editors S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen (Princeton, NJ: Princeton University Press), 362–472.

Liu, M., and Lin, T. I. (2014). A skew-normal mixture regression model. *Educ. Psychol. Meas.* 74, 139–162. doi:10.1177/0013164413498603

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

Lubke, G. H., and Campbell, I. (2016). Inference based on the best-fitting model can contribute to the replication crisis: assessing model selection uncertainty using a bootstrap approach. *Struct. Equ. Model.* 23, 479–490. doi:10.1080/10705511.2016.1141355

Lubke, G. H., Campbell, I., McArtor, D., Miller, P., Luningham, J., and van den Berg, S. M. (2017). Assessing model selection uncertainty using a bootstrap approach: an update. *Struct. Equ. Model.* 24, 230–245. doi:10.1080/10705511.2016.1252265

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

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India.* 2, 49–55.

Masyn, K. (2013). "Latent class analysis and finite mixture modeling," in *The Oxford handbook of quantitative methods in psychology*. Editor T. D. Little (New York, NY: Oxford University Press), Vol. 2, 551–611.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., and Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Dev. Psychol.* 38, 115–142. doi:10.1037/0012-1649.38.1.115

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *J. R. Stat. Soc. Ser. C* 36, 318–324. doi:10.2307/2347790

McLachlan, G., and Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons, 419.

Muthén, B., and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55, 463–469. doi:10.1111/j.0006-341X.1999.00463.x

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol. Methods.* 4, 139–157. doi:10.1037/1082-989X.4.2.139

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equ. Model.* 14, 535–569. doi:10.1080/10705510701575396

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., and Chen, F. (2001). Monte Carlo experiments: design and implementation. *Struct. Equ. Model.* 8, 287–312. doi:10.1207/S15328007SEM0802_7

Peugh, J., and Fan, X. (2015). Enumeration index performance in generalized growth mixture models: a Monte Carlo test of Muthén's (2003) hypothesis. *Struct. Equ. Model.* 22, 115–131. doi:10.1080/10705511.2014.919823

Peugh, J., and Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? A simulation study examining GMM's performance characteristics. *Struct. Equ. Model.* 19, 204–226. doi:10.1080/10705511.2012.659618

Peugh, J., and Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: a Monte Carlo simulation. *Struct. Equ. Model.* 20, 616–639. doi:10.1080/10705511.2013.824780

Rindskopf, D. (2003). Mixture or homogeneous? Comment on Bauer and Curran (2003). *Psychol. Methods.* 8, 364–368. doi:10.1037/1082-989X.8.3.364

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.

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

Serang, S., Jacobucci, R., Stegmann, G., Brandmaier, A. M., Culianos, D., and Grimm, K. J. (2020). Mplus trees: structural equation model trees using mplus. *Struct. Equ. Model.* doi:10.1080/10705511.2020.1726179

Steele, R. J., and Raftery, A. E. (2010). "Performance of Bayesian model selection criteria for Gaussian mixture models," in *Frontiers of statistical decision making and Bayesian analysis*. Editors M.-H. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye (New York, NY: Springer), 113–130.

Stegmann, G., Jacobucci, R., Serang, S., and Grimm, K. J. (2018). Recursive partitioning with nonlinear models of change. *Multivar. Behav. Res.* 53, 559–570. doi:10.1080/00273171.2018.1461602

Tofighi, D., and Enders, C. K. (2008). "Identifying the correct number of classes in growth mixture models," in *Advances in latent variable mixture models*. Editors G. R. Hancock and K. M. Samuelsen (Charlotte, NC: Information Age), 317–341.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333. doi:10.2307/1912557

Wang, K., Yau, K. K., and Lee, A. H. (2002). A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay. *Stat. Med.* 21, 3639–3654. doi:10.1002/sim.1307

Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Comput. Stat. Data Anal.* 50, 1090–1104. doi:10.1016/j.csda.2004.11.004

Yang, C. C., and Yang, C. C. (2007). Separating latent classes by information criteria. *J. Classif.* 24, 183–203. doi:10.1007/s00357-007-0010-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.