



A Comprehensive Evaluation of Model Selection Indices for Class Enumeration in Bayesian Latent Growth Mixture Models

Sarah Depaoli, Ihnwhi Heo, Madelin Jauregui, Haiyan Liu & Fan Jia

To cite this article: Sarah Depaoli, Ihnwhi Heo, Madelin Jauregui, Haiyan Liu & Fan Jia (21 Oct 2025): A Comprehensive Evaluation of Model Selection Indices for Class Enumeration in Bayesian Latent Growth Mixture Models, Structural Equation Modeling: A Multidisciplinary Journal, DOI: [10.1080/10705511.2025.2566135](https://doi.org/10.1080/10705511.2025.2566135)

To link to this article: <https://doi.org/10.1080/10705511.2025.2566135>



© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC



Published online: 21 Oct 2025.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

A Comprehensive Evaluation of Model Selection Indices for Class Enumeration in Bayesian Latent Growth Mixture Models

Sarah Depaoli , Ihnwhi Heo , Madelin Jauregui , Haiyan Liu , and Fan Jia 

University of California, Merced

ABSTRACT

Class enumeration remains one of the most critical and error-prone steps in latent growth mixture modeling (LGMM), particularly within the Bayesian framework. This study provides a comprehensive simulation-based evaluation of Bayesian model selection indices, focusing on the impact of likelihood formulation (marginal used in this case) and Dirichlet prior specification for class proportions. Although Bayesian methods offer flexibility and robustness in estimating complex models, missteps in class enumeration or inappropriate prior specification can bias results, mislead substantive conclusions, and impair model fit. We systematically varied true population structures and prior specifications to assess how these factors interact to affect model selection accuracy across various indices. We examined the performance of several Bayesian indices: the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC), the leave-one-out information criterion (LOOIC), the expected Akaike information criterion (EAIC), and the expected Bayesian information criterion (EBIC). Our study contributes practical recommendations for researchers conducting Bayesian LGMM, highlighting methodological best practices and key areas for further development with respect to model comparison and selection indices in the Bayesian framework. These results advance our understanding of model selection behavior in complex Bayesian mixture models and provide a foundation for improving estimation and inference in applied research.

KEYWORDS



Bayesian estimation; class enumeration; Dirichlet prior; growth mixture modeling; prior alignment

Model misspecification within structural equation modeling (SEM) remains a critical and persistent area of methodological research (e.g., Cao & Liang, 2022; Depaoli et al., 2024; Liu, Heo, Depaoli, et al., 2025; McNeish & Harring, 2017; West et al., 2012). Numerous specification challenges, including errors in both measurement and structural components, can compromise research quality, undermine the accuracy of parameter estimation and model selection, and ultimately lead to misleading substantive research conclusions (Cain & Zhang, 2019; Depaoli et al., 2023, 2024; Heo et al., 2024; Liu, Heo, Depaoli, et al., 2025; Liu, Heo, Ivanov, et al., 2025; McNeish & Harring, 2017; Winter & Depaoli, 2022). These issues become even more complex and pose substantial challenges to researchers when SEM incorporates finite mixture modeling to address unobserved latent classes and thus account for population heterogeneity. In such contexts, the risk of misspecification expands beyond ensuring proper measurement and structural specification. Researchers must carefully consider how to estimate and select the appropriate latent class structure.

One widely used modeling framework that embodies both SEM and mixture modeling is the latent growth mixture model (LGMM; B. Muthén, 2001; B. Muthén et al., 1998), which combines longitudinal data analysis with class-based heterogeneity to uncover distinct latent trajectory classes over time. Within the LGMM framework, one of the most central

challenges is class enumeration—identifying, specifying, and selecting the optimal number of latent classes. Class enumeration errors can manifest in either overextraction or underextraction of latent classes, resulting in convergence problems, estimation bias, and misclassification. From a frequentist perspective, these methodological issues have been extensively documented (McNeish, 2023; McNeish & Harring, 2017; Nylund-Gibson & Choi, 2018; Tueller & Lubke, 2010), and various model selection indices have been evaluated for their effectiveness in detecting class structure misspecification (Nylund et al., 2007). As such, model specification in the finite mixture modeling framework should additionally regard the reliability and validity of latent class solutions. The ability to properly detect model class structures is directly intertwined with the ability to correctly interpret latent group differences and generalize findings.

Over the past two decades, following Nylund et al.'s foundational contribution, significant advancements in estimation techniques have extended into Bayesian frameworks. Bayesian estimation has gained increasing popularity in SEM (i.e., Bayesian SEM) and mixture modeling due to its flexibility, its capability to incorporate prior information, and its robust handling of model complexity (Depaoli, 2013, 2014; S.-Y. Kim, 2014; S.-Y. Kim et al., 2013; S. Kim et al., 2022; Kohli et al., 2015; Tong et al.,

CONTACT Sarah Depaoli  sdepaoli@ucmerced.edu  Department of Psychological Sciences, University of California, Merced, 5200 N. Lake Road, Merced, CA 95343, USA.

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

2022; Tong & Ke, 2016, 2021; Tong & Zhang, 2020). However, the implementation of Bayesian LGMM introduces two underappreciated yet critical sources of model vulnerability. First, model misspecification due to incorrect class enumeration persists in the Bayesian context, often compounded by the choice of fit indices and how the likelihood of those indices is formulated (Merkle et al., 2019; Tong et al., 2022). Second, the specification of prior distributions, particularly those governing class proportions via the Dirichlet prior, introduces an additional layer of complexity. When these priors are misaligned with the true population structure (e.g., assuming equal class proportions when the population is highly imbalanced), they can bias class assignment, model fit, and selection outcomes (Depaoli, 2013, 2014; Depaoli et al., 2017; S. Kim et al., 2021; Tong et al., 2022).

Prior research has begun to address the first issue. For instance, Tong et al. (2022) examined how likelihood formulation (marginal vs. conditional) affects the performance of Bayesian model selection indices in the context of class enumeration in LGMMs. Their findings indicated that marginal likelihood-based model selection indices generally outperform their conditional counterparts, which resonated with similar findings found in other non-mixture SEM models (Merkle et al., 2019). However, this topic still calls for extensive simulation. Moreover, the influence of prior specifications, particularly how priors on class proportion and population characteristics affect the performance of model selection indices, has not yet been considered, an issue we turn to next.

Critically, the second issue—prior (mis)alignment—remains virtually unexplored in the Bayesian LGMM literature. Although the use of Dirichlet priors is widespread (Depaoli, 2013; Depaoli et al., 2017; Tong et al., 2022; Van Erp et al., 2018; Yang & Dunson, 2010), little is known about how mismatches between prior assumptions and true population class proportions influence model performance. To illustrate, when the population includes a minority class but the prior assumes equal class sizes, the resulting posterior estimates may underrepresent that minority class and lead to underextraction of latent classes. In a similar vein, using highly informative but incorrect priors for class proportions can distort model fit assessments and compromise class recovery. The reality is that there will likely be some degree of mismatch with the prior in applied settings, since the true population-level proportions are unknown. Despite its theoretical and practical relevance, this issue of prior misalignment has received minimal attention in methodological research.

The present study addresses both of these methodological challenges in tandem. Our investigation evaluates the performance of a broad set of Bayesian model selection indices, all computed using marginal likelihoods—a choice we justify based on its conceptual consistency with traditional SEM practices (as elaborated in subsequent sections) and its favorable performance demonstrated in findings by Tong et al. (2022); Merkle et al., 2019; Du et al. (2024). Our particular focus is on examining the consistency and

accuracy of these indices in detecting latent class structure misspecification under different scenarios of prior settings. To this end, we conduct a comprehensive simulation study that systematically manipulates design factors, including (1) model (mis)specification in class enumeration and (2) (mis)alignment between population-level class proportions and those specified via the Dirichlet prior. Our simulation design improves upon prior work by systematically investigating how model misspecification and prior misalignment interact to affect model convergence and the selection of the true latent class solution. To our knowledge, this is the first study to explicitly assess the impact of prior misalignment on class enumeration in Bayesian LGMMs. Our hope is that the current findings will help researchers make more informed decisions regarding class enumeration, latent class solution selection, and the thoughtful specification of priors in applied settings—recognizing that priors may not always align perfectly with population characteristics and thus require practical, data-informed guidance.

1. Organization of the Current Investigation

This paper is organized as follows. We first begin with an overview of the benefits of Bayesian estimation for latent variable models, and then extend this discussion to the LGMM, which is the model we focus on here. We then present the formulation and notation for the LGMM, including the relevant prior distributions. Next, we turn our attention to the methods that are currently available for detecting model (mis)fit in Bayesian latent variable modeling. We present notation and descriptions for the most commonly implemented indices, which are a major focus in the current investigation. We then tie these topics together by presenting rationale for further exploration regarding the use of these indices for detecting model misfit in the LGMM. That links directly to the simulation design and results, which are presented next, including a secondary simulation examining an extreme yet realistic research scenario. We conclude the paper with a discussion of our findings, points that applied researchers should consider when implementing these indices in practice, and recommendations for future methodological developments regarding fit and assessment measures in the Bayesian estimation framework.

2. Benefits of Bayesian Estimation for Latent Variable Models

Several seminal papers have been written about the general benefits of Bayesian methods (see, e.g., Carlin & Louis, 2000; Gelman et al., 2014; B. O. Muthén & Asparouhov, 2012; van de Schoot et al., 2017; 2021). In addition, the popularity of this framework has been steadily on the rise, especially for SEM (van de Schoot et al., 2017). There are many potential reasons for increased use and exposure to these methods with SEM, and we briefly highlight the most relevant reasons here.

Within SEM, model complexity is tied to accurate parameter recovery and convergence issues (S.-Y. Kim et al., 2013). More complex models can produce problems with convergence, and there can also be issues with obtaining inaccurate parameter estimates (sometimes due to the non-convergence issue and sometimes not). Bayesian methods have been shown to aid in solving these issues in a variety of different SEM model-forms (see, e.g., S.-Y. Kim et al., 2013). In addition, SEM has been discussed in a much more flexible manner in the Bayesian framework (B. O. Muthén & Asparouhov, 2012), where processes requiring strict model constraints (e.g., model invariance testing) can be carried out in a more flexible, or approximate, manner through the implementation of priors.

One area where Bayesian methods have shown to be of particular benefit is in the estimation of latent mixture models. Specifically, mixture (or latent class) models carry an added complexity of estimating the class structure. Researchers often rely on substantive knowledge and a collection of model fit or assessment measures to help determine the number of latent classes from a set of possible solutions. Previous simulation research by Depaoli (2013) has shown that, even when estimating the correct number of latent classes, it can be difficult to properly estimate the *size* of those latent classes via class proportion estimates. That estimation accuracy issue is tied to factors such as the number of classes, whether there is a strong majority class (i.e., a class with a much larger proportion of cases assigned) or a minority class, and also separation—the concept of how overlapping, or distinctive, latent classes are from one another. Especially when class proportions are quite different across latent groups (e.g., there is a very large or very small class), and separation is more difficult to distinguish, the Bayesian framework has been shown to greatly enhance the accuracy of results (Depaoli, 2013, 2014; Depaoli et al., 2017; S. Kim et al., 2021; Tong et al., 2022). The use of priors, particularly for the latent class proportions, appears to benefit the results beyond what can be produced using conventional estimation techniques. However, the selection and implementation of these priors should be done with intent and transparency, as even a slight modification of the hyperparameters can alter the findings (Depaoli et al., 2017).

3. The Latent Growth Mixture Model: Notation and Priors

The LGMM can be used for tracking change patterns over time, and the mixture component acts as an extension to the simpler latent growth curve model (which does not include latent classes). This section borrows notation detailed in Depaoli (2021). For the LGMM, the data are assumed to have been generated from a mixture distribution, where there are $c = 1, 2, \dots, C$ latent classes of proportion π_c , each allowed its own set of parameters as detailed in the model equations. The model can be separated into a measurement and a structural part of the model. The measurement part of the model can be written as:

$$y_{ic} = \Lambda_y \eta_{ic} + \epsilon_{ic}, \quad (1)$$

where y_{ic} is a vector of observed repeated-measure data for person i in latent class c , Λ_y is a $T \times m$ matrix of factor loadings (T = number of time points; m = number of latent factors). Column 1 in Λ_y is fixed to 1's, and the remaining $m - 1$ columns contain information about the time scale and slope shape for data collection (e.g., 0, 1, 2, 3 for four equally spaced time points and a linear slope). The vector, η_{ic} , contains the m latent growth parameters (e.g., intercept and slope), and ϵ_{ic} is a vector of normally distributed measurement errors (assumed centered at zero).

The structural part of the model is as follows:

$$\eta_{ic} = \alpha_c + \zeta_{ic}, \quad (2)$$

where vector η_{ic} still contains the growth parameters, α_c is a vector of factor means, and ζ_{ic} is a vector of normally distributed (centered at zero) deviations of parameters from their population means. The reduced form of the equation is:

$$y_{ic} = \Lambda_y(\alpha_c + \zeta_{ic}) + \epsilon_{ic}. \quad (3)$$

From this formulation, the model-implied mean and covariance can be respectively written as follows:

$$\mu_c(\theta) = \Lambda_y \alpha_c, \quad (4)$$

$$\Sigma_c(\theta) = \Lambda_y \Psi_\eta \Lambda_y' + \Theta_{\epsilon_c}, \quad (5)$$

where ζ_{ic} can be omitted, as the expectation of η is equal to α . Here, $\mu_c(\theta)$ is the mean vector of the y 's, and $\Sigma_c(\theta)$ is the covariance matrix of the y 's. Further, Ψ_η is the latent factor covariance matrix, and the covariance matrix for the manifest variable errors is Θ_{ϵ_c} . In this expression, the latent factor covariance matrix does not contain a c subscript, which indicates homogeneity across classes, but this can be relaxed by adding a c subscript. Figure 1 shows a diagram of the basic form of the LGMM.

The LGMM can be implemented in the frequentist or Bayesian estimation frameworks. For Bayesian estimation, model priors must be defined for each parameter in the model. The most common model priors for the main elements of the LGMM are as follows:

$$\begin{aligned} \pi &\sim \mathcal{D}[d_1 \dots d_C], & \alpha_{mc} &\sim \mathcal{N}[\mu_{\alpha_{mc}}, \sigma_{\alpha_{mc}}^2], \\ \theta_{\epsilon_{err}} &\sim \mathcal{IG}[a_{\theta_{err}}, b_{\theta_{err}}], & \Psi_\eta &\sim \mathcal{IW}[\Psi, \nu]. \end{aligned}$$

The latent class proportion for each class is denoted π_c , and the vector of class proportions for all C classes (π) is typically modeled using a Dirichlet (\mathcal{D}) distribution. The Dirichlet distribution hyperparameters (d_1, \dots, d_C) represent hyperparameters reflecting the class sizes. Depending on the software implemented, these hyperparameters may be formed in terms of proportions, number of people, or thresholds. An extensive discussion of these differences is provided in Chapter 10 of Depaoli (2021). The growth factor means (α) are typically assumed normally distributed (\mathcal{N}). Here, α_{mc} is the latent factor mean for factor $m = 1, \dots, M$ and latent class $c = 1, \dots, C$. The two hyperparameters are $\mu_{\alpha_{mc}}$ (expectation for the factor mean; mean hyperparameter) and $\sigma_{\alpha_{mc}}^2$ (variance hyperparameter).

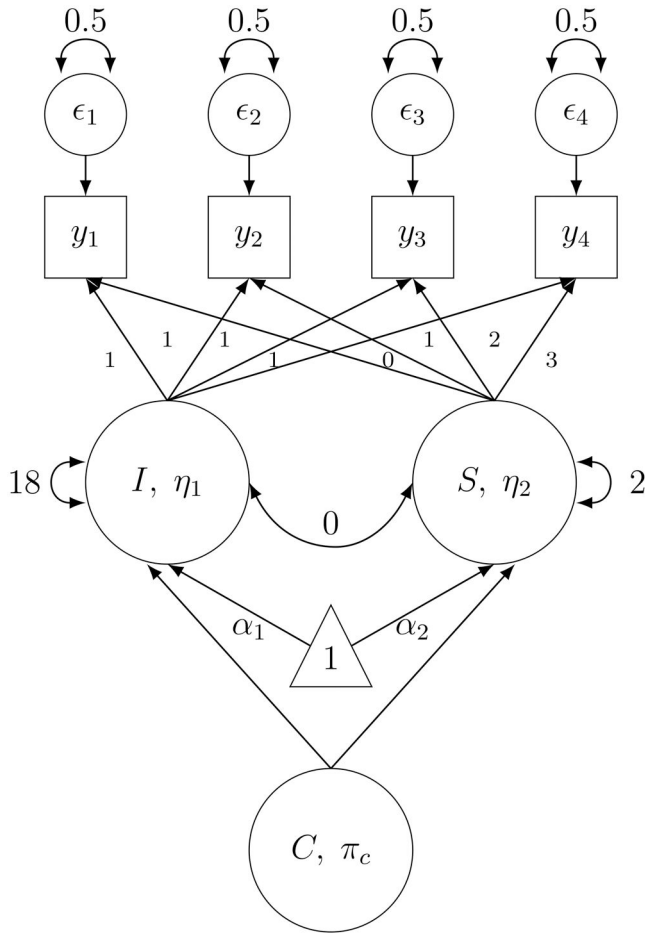


Figure 1. Latent growth mixture model.

Next, Θ_{ec} (error variance matrix) can be linked to a prior. When the error variances are assumed uncorrelated, then univariate priors can be placed on the individual elements in Θ_{ec} , denoted as θ_{err} (representing an individual element in the $r \times r$ matrix). The prior for this element is typically defined by an inverse gamma (\mathcal{IG}) distribution, with hyperparameters representing the shape ($a_{\theta_{err}}$) and scale ($b_{\theta_{err}}$) of the distribution. Finally, Ψ_η is the latent factor covariance matrix and it can receive the inverse Wishart (\mathcal{IW}) prior distribution, with hyperparameters representing a positive definite matrix of size p (Ψ) and degrees of freedom (ν). Just as with any prior, this one can be set to vary across latent classes if desired. In all cases, manipulating the hyperparameters controls the level of (un)certainty or informativeness in the corresponding prior distribution.

4. Information Criteria in the Bayesian Framework

The current investigation examines the ability of a variety of Bayesian information criteria as model selection indices to identify proper class enumeration (e.g., correct number of latent classes to reflect the population). We included several indices here to provide a full landscape and comparison for commonly implemented tools. Next, we present details for the deviance information criterion (DIC; Spiegelhalter et al., 2002), the Watanabe-Akaike information criterion (WAIC;

Watanabe, 2010), the leave-one-out information criterion (LOOIC; Vehtari et al., 2017), the expected Akaike information criterion (EAIC; Carlin & Louis, 2000) and the expected Bayesian information criterion (EBIC; Carlin & Louis, 2000), where the EAIC and EBIC are a Bayesian analog of the Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978), and incorporates the expectation over the posterior distribution of model parameters (Carlin & Louis, 2000; Spiegelhalter et al., 2002). In all cases, the information criteria are interpreted in the same way. Specifically, lower estimates correspond with the optimal model when comparing information criteria values across several competing models.

In this study, the information criterion is computed based on the marginal likelihood. For the LGMM, the marginal likelihood is defined as

$$L(\boldsymbol{\pi}, \boldsymbol{\theta} | \mathbf{y}) = \prod_{i=1}^n \sum_{c=1}^C \pi_c p(\mathbf{y}_i | \boldsymbol{\mu}_c(\boldsymbol{\theta}), \boldsymbol{\Sigma}_c(\boldsymbol{\theta})),$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_C)$ is the vector of class proportions, constrained such that $\sum_{c=1}^C \pi_c = 1$. The parameter vector $\boldsymbol{\theta}$ includes all the model parameters, such as the class-specific mean $\boldsymbol{\alpha}_c$ and the covariance matrices $\boldsymbol{\Psi}_c$ of the latent intercept and slope, and the residual variances Θ_c . The function $p(\mathbf{y}_i | \boldsymbol{\mu}_c(\boldsymbol{\theta}), \boldsymbol{\Sigma}_c(\boldsymbol{\theta}))$ denotes the likelihood of observation \mathbf{y}_i given the model implied means $\boldsymbol{\mu}_c(\boldsymbol{\theta})$ and covariance matrix $\boldsymbol{\Sigma}_c(\boldsymbol{\theta})$ of class c .

Given a set of posterior samples $(\boldsymbol{\pi}^1, \boldsymbol{\theta}^1)$, $(\boldsymbol{\pi}^2, \boldsymbol{\theta}^2), \dots, (\boldsymbol{\pi}^S, \boldsymbol{\theta}^S)$ for model parameters, we would obtain a sample for the log-likelihood D with $D^s(\boldsymbol{\pi}, \boldsymbol{\theta}) = D(\boldsymbol{\pi}^s, \boldsymbol{\theta}^s)$,

$$D^s(\boldsymbol{\pi}, \boldsymbol{\theta}) = -2 \log(L(\boldsymbol{\pi}^s, \boldsymbol{\theta}^s | \mathbf{y})).$$

4.1. DIC

The DIC was proposed by Spiegelhalter et al. (2002) to evaluate the fit of the Bayesian model while considering the complexity of the model, despite some arguments on its robustness (Spiegelhalter et al., 2014). The DIC can be written in the following two equivalent forms with $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})$ being the posterior mean of the model parameters,

$$\text{DIC} = \overline{D(\boldsymbol{\pi}, \boldsymbol{\theta})} + p_D \text{ or } \text{DIC} = D(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}) + 2p_D,$$

with p_D representing the complexity of the model, which is defined as the discrepancy between the mean of the deviance $\overline{D(\boldsymbol{\pi}, \boldsymbol{\theta})}$ and deviance evaluated at the posterior mean of the model parameters, $p_D = \overline{D(\boldsymbol{\pi}, \boldsymbol{\theta})} - D(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}})$.

4.2. WAIC

The WAIC was proposed by Watanabe (2010) and is computed based on the log-pointwise predictive density. WAIC offers several advantages, including being invariant to the reparameterization and applicability to singular models where DIC may fail (Gelman et al., 2014; Vehtari et al., 2017).

For a mixture model, the WAIC is defined as:

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}, \quad (6)$$

where the log-pointwise predictive density (lppd) is

$$\text{lppd} = \sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S p(y_i | \pi^s, \theta^s) \right), \quad (7)$$

where S is the number of posterior samples, and (π^s, θ^s) represents the s -th posterior draw. The effective number of parameters, p_{WAIC} , serves as the penalty of a model complexity and is calculated as the sum of posterior variances of the log-likelihood across data points:

$$p_{\text{WAIC}} = \sum_{i=1}^n \frac{1}{S} \sum_{s=1}^S (\log p(y_i | \pi^s, \theta^s) - \frac{1}{S} \sum_{s=1}^S \log p(y_i | \pi^s, \theta^s))^2.$$

Thus, the WAIC provides an estimate of out-of-sample predictive accuracy while adjusting for model complexity.

4.3. LOOIC

The LOOIC is a Bayesian model selection index based on leave-one-out cross-validation (Vehtari et al., 2017). It provides an estimate of out-of-sample predictive accuracy and is computed as:

$$\text{LOOIC} = -2 \widehat{\text{elpd}}_{\text{loo}}, \quad (8)$$

where the expected log pointwise predictive density (elpd) under leave-one-out cross-validation is defined as

$$\widehat{\text{elpd}}_{\text{loo}} = \sum_{i=1}^n \log p(y_i | y_{-i}) = \sum_{i=1}^n \log \left(\int p(y_i | \theta) p(\theta | y_{-i}) d\theta \right),$$

where $p(y_i | y_{-i})$ denotes the leave-one-out predictive density for observation y_i , given the dataset excluding the i th observation.

In practice, this quantity is estimated using Pareto-smoothed importance sampling (PSIS; Vehtari et al., 2017). The PSIS-based estimate of the expected log-pointwise predictive density is given by:

$$\widehat{\text{elpd}}_{\text{psis-loo}} = \sum_{i=1}^n \log \left(\frac{\sum_{s=1}^S \omega_i^s p(y_i | \theta^s)}{\sum_{s=1}^S \omega_i^s} \right),$$

where ω_i^s are importance sampling weights. This estimation is implemented in the `loo` package in R (Vehtari et al., 2017).

4.4. EAIC

The EAIC is an extension of the AIC that incorporates posterior expectations, making it suitable for use in a Bayesian framework (e.g., Carlin & Louis, 2000). Like the AIC, the EAIC balances model fit and complexity but is computed using Bayesian posterior samples rather than point

estimates. Given the posterior samples (π^1, θ^1) , (π^2, θ^2) , ..., (π^S, θ^S) , the sample-level EAIC is computed as:

$$\text{EAIC} = -2 \sum_{i=1}^n \left(\frac{1}{S} \sum_{s=1}^S \log p(y_i | \pi^s, \theta^s) \right) + 2 p_{\text{EAIC}},$$

where the effective number of parameters is given by:

$$p_{\text{EAIC}} = \sum_{i=1}^n \left[\frac{1}{S} \sum_{s=1}^S \log p(y_i | \pi^s, \theta^s) - \log p(y_i | \bar{\pi}, \bar{\theta}) \right],$$

and $\bar{\pi}$ and $\bar{\theta}$ denote the posterior means of the parameters.

4.5. EBIC

The EBIC (Carlin & Louis, 2000) is conceptually similar to the EAIC, but it employs a different penalty term that grows with sample size, following the spirit of the BIC. Both criteria use posterior samples to account for model uncertainty.

The sample-level EBIC is defined as:

$$\text{EBIC} = -2 \sum_{i=1}^N \left(\frac{1}{S} \sum_{s=1}^S \log p(y_i | \pi^s, \theta^s) \right) + \log(n) \cdot p_{\text{EBIC}},$$

where the effective number of parameters is given by:

$$p_{\text{EBIC}} = \sum_{i=1}^n \left[\frac{1}{S} \sum_{s=1}^S \log p(y_i | \pi^s, \theta^s) - \log p(y_i | \hat{\pi}, \hat{\theta}) \right],$$

and $\hat{\pi}$ and $\hat{\theta}$ represent the posterior means of the model parameters.

4.6. Marginal vs. Conditional Likelihoods for Information Criteria

In the current study, the information criteria are computed based on the marginal likelihood, integrating over the random components, including the latent growth factors and class assignments. We acknowledge that these fit indices can, in general, be defined for any form of the likelihood, including the conditional likelihood given the latent growth factors and class assignments (Celeux et al., 2006; Gelman et al., 2014; Spiegelhalter et al., 2002).

In existing software for Bayesian SEM, such as JAGS, BUGS, or Stan (Carpenter et al., 2017; D. J. Lunn et al., 2000; D. Lunn et al., 2012; Plummer, 2003), the deviance is typically computed based on the conditional likelihood (Merkle et al., 2019). However, treating latent factor scores and class memberships as actual parameters of interest can lead to substantially different assessments of model complexity. In the general SEM framework, latent factor scores are not considered model parameters. Consequently, the fit indices in the frequentist SEM tradition are all based on the marginal likelihood, where the latent variables are integrated.

To maintain consistency with the frequentist framework and likelihood-based fit indices, we computed the information criteria based on the marginal likelihood. This approach aligns with recent discussions in the literature. For example, Tong et al. (2022) investigated the impact of using

marginal versus conditional likelihoods on class enumeration in Bayesian LGMMs and found that marginal likelihood-based indices such as the DIC, WAIC, and leave-one-out cross-validation outperformed their conditional counterparts. These findings are consistent with earlier results from non-mixture SEM models (Du et al., 2024; Merkle et al., 2019), which also emphasized the limitations of conditional likelihood-based comparisons. This distinction is important, as it can substantially influence model selection outcomes. In line with these recommendations, our implementation relies on marginal likelihoods for all model selection index computations to maintain conceptual coherence with the SEM tradition.

5. Brief Literature Recap

Previous literature has focused on the performance of various information criteria in terms of model misspecification and class enumeration. Specifically, the majority of the Bayesian work has focused on the DIC, BIC, EAIC, EBIC, and others in non-mixture models, such as confirmatory factor analysis (CFA), SEM, latent growth curve modeling (LGCM), and item response theory (IRT). Key findings reveal nuanced performance variations across different model specifications, prior selections, and sample sizes.

For example, Cain and Zhang noted that within SEM, the true model detection rates of the DIC against underfitting models improved as the sample size, model size, and degree of model misspecification increased. Informative priors were found to be superior to diffuse priors, although the influence of priors decreased as sample size increased. Related to the role of priors, Liu et al. (2022) examined the impact of priors on different locations and revealed that the DIC is more sensitive when selecting the true model, which was more complex than the misspecified (underfitting) model. Moreover, several studies in LGCM compared the performance of the BIC and DIC in various forms of misspecification (Depaoli et al., 2023, 2024; Heo et al., 2024; Winter & Depaoli, 2022). Winter and Depaoli highlighted that with a quadratic latent growth model, the BIC, and to a lesser extent the DIC (less sensitive to sample size), preferred more parsimonious models over the true model, while both could correctly identify the true model compared with overfitting models. Depaoli et al. (2023) and Heo et al. (2024) reached similar conclusions in the context of the piecewise growth model. Specifically, they found that the DIC outperformed the BIC when detecting misplacement or ignorance of the change point, while both required large sample sizes. Conversely, in the context of CFA, Depaoli et al. (2024) found that the BIC consistently outperformed the DIC in model selection, especially in overfitting scenarios. These seemingly contradictory findings across studies reflect the inherent complexity of evaluating information criteria performance in Bayesian modeling, as results can be impacted by many factors, such as the selection of priors, the nature of model misspecification, and sample size. Notably, a consistent pattern emerging from these studies was that the BIC tended to prefer

parsimonious models compared to the DIC. The specification of the DIC may also play a role in performance, with Du et al. (2024) indicating that the marginal-likelihood-based version outperformed the conditional-likelihood-based version.

Regarding the LGMMs, in particular, studies on the performance of information criteria for identifying the correct number of classes have been found in the frequentist framework. For example, Nylund et al. (2007) studied class enumeration in a linear LGMM with two classes and noted that the BIC generally worked well and outperformed the AIC in identifying the correct number of classes, although it was sensitive to small sample sizes. In contrast, other studies revealed that both the AIC and BIC could perform poorly with highly complex models, such as multiple-class models with heterogeneous growth patterns (e.g., Peugh & Fan, 2012; Tofighi & Enders, 2008) and multiphase LGMMs (e.g., S.-Y. Kim, 2014). Also regarding the LGMM, S. Kim et al. (2021) explored index performance in the context of different variations of the model, including the conventional LGMM formulation, *t*-based version (allowing for thicker tails), and the median-based formulation of the model. The DIC, WAIC, and LOO-CV were compared across these model types. The authors found that proper model selection was most consistent for the *t*-based and median-based formulations of the LGMM. The conventional specification of the LGMM was linked to poorer selection accuracy among these indices, especially when outliers were present. Overall, the literature on class enumeration demonstrates that the performance of information criteria is influenced by multiple interacting factors, including class separation, class proportion, model complexity, sample size, and even estimation procedures (McNeish & Harring, 2017).

The performance of additional Bayesian model selection indices, such as the EAIC and EBIC, has previously been evaluated in the context of IRT models. According to Bolfarine and Bazan (2010), these two indices, along with the DIC, demonstrated good performance in estimating ability parameters, with the DIC and EAIC performing similarly in favoring skewed logistic IRT models. In a later study, da Silva et al. (2019) found that the performance of the EAIC and EBIC was sensitive to both sample size and the number of items. As a result, the authors advised against using the EAIC and EBIC in models with small numbers of respondents or items. The performance of several common Bayesian and non-Bayesian indices were also studied in the IRT-framework. In particular, Luo and Al-Harbi (2017) considered the fully Bayesian indices of WAIC and LOO-CV and found that these indices performed better than the conventional methods of the likelihood ratio test, AIC, BIC, and DIC. They further highlighted the inconsistencies in the AIC performance for proper model detection under different conditions such as sample size and test length. In addition, Fujimoto and Falk (2024) examined the DIC, WAIC, and LOO-CV in the context of multidimensional IRT. The general findings suggested that the DIC favored certain IRT models over others, even when they represented model mis-

specifications. The DIC showed much more bias and patterns of incorrect model selection as compared to the WAIC and LOO-CV.

Despite these investigations across various modeling contexts, an important gap remains in the literature: No study has yet examined the performance of a broad range of information criteria under both class enumeration and prior misalignment within Bayesian mixture models. Since the Bayesian indices we covered above may be helpful tools for identifying class structures, our goal is to assess this capability in terms of prior (mis)alignment. In the following sections, we detail a comprehensive simulation study aimed at uncovering this methodological complexity within the Bayesian framework.

6. Simulation Design

This study uses a simulation design to evaluate the performance of several marginal likelihood-based information criteria (DIC, WAIC, LOOIC, EAIC, and EBIC) in distinguishing correct from incorrect model specifications in LGMMs, under conditions where the priors for true class proportions are either aligned or misaligned. The performance of each criterion was assessed across scenarios that varied three key factors: sample size (150, 300, 900), class separation as indexed by Mahalanobis distance (2.7, 3.2, 3.7), and latent class proportions (equal vs. unequal). The analysis model specification was manipulated along two dimensions: the number of latent classes (1-class, 2-class, 3-class, and 4-class solutions) and the type of prior distributions specified for class proportions. For 1-class solutions, only a diffuse prior was used, yielding $3 \text{ (sample sizes)} \times 3 \text{ (class separation levels)} \times 2 \text{ (class proportions)} \times 1 \text{ (prior)} = 18$ conditions. For 2-class, 3-class, and 4-class solutions, three types of priors were examined: a diffuse prior, an informative prior assuming equal class sizes, and an informative prior assuming unequal class sizes, resulting in $3 \times 3 \times 2 \times 3 = 54$ conditions for each class solution. With three such class solutions, this yields $54 \times 3 = 162$ additional conditions. In total, $18 \text{ (from 1-class)} + 162 \text{ (from 2–4-class)} = 180$ unique simulation conditions were generated. Each condition included 500 replications, resulting in 90,000 total simulated datasets.

6.1. Population Model

The population model was a linear LGMM, consisting of four time points and two latent classes. The residual variance was fixed at 0.5, while the variances of the intercept and slope were set to 18 and 2, respectively, with no covariance between the growth factors (Figure 1). Population values for the LGMM were determined based on the degree of class separation (described below; Depaoli, 2013; S.-Y. Kim, 2014; Tong et al., 2022) and are presented in Table 1.

6.2. Sample Size

We designed the study to include three different sample size conditions: a small sample with 150 participants, a medium sample with 300 participants, and a large sample with 900 participants. These varying sample sizes allowed for a more comprehensive assessment of the model's performance across different population sizes (Depaoli, 2013; S. Kim et al., 2022; Tong et al., 2022).

6.3. Class Separation

Growth parameter means were adjusted to represent three levels of class separation, determined by the multivariate Mahalanobis distance (MD) obtained by $\{(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)\}^{1/2}$, where μ_1 and μ_2 denote the mean vectors of the first and second latent classes, respectively, and Σ^{-1} is the inverse of the common covariance matrix. These levels corresponded to a small (MD = 2.7), moderate (MD = 3.2), and large (MD = 3.7) degree of separation between Class 1 and Class 2 (S.-Y. Kim, 2014; S. Kim et al., 2021; Tong et al., 2022). A comprehensive outline of the intercept and slope values for each level of class separation is provided in Table 1. To ensure realistic and interpretable parameter settings, we modified the intercept and slope values for Class 2 based on the specified class separation levels and informed by population values reported in Depaoli (2013).

6.4. Class Proportion

Previous mixture studies indicate that unequal class proportions can influence both the accuracy of model selection

Table 1. Summary of simulation population parameters.

Simulation factor	Levels
Sample size	$n = 150, 300, 900$
Class proportion	Equal (50/50) Unequal (70/30)
Class separation	MD = 2.7 (small), 3.2 (medium), 3.7 (large)
Population intercept and slope	Class 1 $\beta_0 = 48 \quad \beta_1 = 3$ Class 2 $\beta_0 = 40.913 \quad \beta_1 = 0$ (small) $\beta_0 = 37.835 \quad \beta_1 = 0$ (medium) $\beta_0 = 35.138 \quad \beta_1 = 0$ (large)
Model specification	True Model (2 class) Estimated Models: Underspecified (1 class) and overspecified (3 and 4 class)

and parameter recovery (Depaoli, 2013; Tueller & Lubke, 2010). Therefore, we varied the class proportions at the population level, considering both equal (1:1) and unequal (7:3) ratios.

6.5. Model Specification

6.5.1. Class Enumeration

We varied the model specification to include a 1-class (underspecified), 2-class (correctly specified), 3-class, and 4-class (overspecified) solution. This approach enabled us to assess each index's ability to distinguish between correct and incorrect model specifications, including those that under- or overestimate the number of latent classes.

6.5.2. Class Proportion Priors

The Dirichlet prior for class proportions in each solution was set to one of three types: diffuse, informative equal, or informative unequal. Our focus was to study the impact of using correct versus incorrect informative priors in comparison to standard diffuse priors. The class proportion priors varied across different sample sizes and model specifications, as detailed in Table 2. Note that for the underspecified 1-class solution, only diffuse priors were applied.

6.6. Bayesian Estimation

On *Mplus* Version 8.7 (L. K. Muthén & Muthén, 1998–2017), data were generated, and the Bayesian estimation approach was implemented. Each model was estimated using a single chain with 40,000 iterations, with the first half discarded as burn-in. The total number of iterations was selected based on preliminary tests of different chain lengths to ensure adequate convergence while minimizing unnecessary computational burden. To prevent between-chain label switching, only one Markov chain was used. To address within-chain label switching, parameter identifiability constraints were applied (Cassiday et al., 2021). Specifically, for models with two or more classes, constraints were imposed such that the intercept of Class 1 was restricted to be greater than that of Class 2. This ensured a consistent ordering of latent classes across replications.

Table 2. Class proportion priors.

Class solution	Diffuse	Informative equal	Informative unequal
<i>n</i> = 150			
2-class	$\mathcal{D}(10, 10)$	$\mathcal{D}(75, 75)$	$\mathcal{D}(105, 45)$
3-class	$\mathcal{D}(10, 10, 10)$	$\mathcal{D}(50, 50, 50)$	$\mathcal{D}(105, 30, 15)$
4-class	$\mathcal{D}(10, 10, 10, 10)$	$\mathcal{D}(37.5, 37.5, 37.5, 37.5)$	$\mathcal{D}(105, 15, 15, 15)$
<i>n</i> = 300			
2-class	$\mathcal{D}(10, 10)$	$\mathcal{D}(150, 150)$	$\mathcal{D}(210, 90)$
3-class	$\mathcal{D}(10, 10, 10)$	$\mathcal{D}(100, 100, 100)$	$\mathcal{D}(210, 60, 30)$
4-class	$\mathcal{D}(10, 10, 10, 10)$	$\mathcal{D}(75, 75, 75, 75)$	$\mathcal{D}(210, 30, 30, 30)$
<i>n</i> = 900			
2-class	$\mathcal{D}(10, 10)$	$\mathcal{D}(450, 450)$	$\mathcal{D}(630, 270)$
3-class	$\mathcal{D}(10, 10, 10)$	$\mathcal{D}(300, 300, 300)$	$\mathcal{D}(630, 180, 90)$
4-class	$\mathcal{D}(10, 10, 10, 10)$	$\mathcal{D}(225, 225, 225, 225)$	$\mathcal{D}(630, 90, 90, 90)$

Note. The true number of classes was two, with 1:1 class proportions for equal conditions and 7:3 for unequal conditions.

To mitigate issues related to parameter solutions reaching local maxima, we tested a range of custom random start values for each condition. These included sets perturbed from the true parameter values and purely random initializations. These strategies were used to improve the likelihood of convergence to the global maximum and to reduce the risk of local solutions. For the 2-class solution, the true parameter values were used as starting values. For the other solutions (1-class, 3-class, and 4-class), we specified custom initial values tailored to each model. These starting values proved effective, consistently avoiding convergence to local solutions across all replications.

Following model estimation across simulation conditions, posterior chains were extracted from *Mplus* and imported into R for the computation of marginal likelihood-based indices. Specifically, we calculated the log-likelihoods and implemented the DIC, EAIC, and EBIC directly in R, while the LOOIC and WAIC were computed using the *loo* package (Vehtari et al., 2017). The log-likelihood values were extracted from the posterior samples, so that pointwise log-likelihoods for each observation and posterior draw were retained, as required for the computation of the LOOIC and WAIC.

6.7. Outcomes of Interest

Our evaluation focuses on two primary outcomes: convergence rates and selection rates. Convergence rates reflect the stability and computational feasibility of the models under different simulation conditions, whereas selection rates assess the accuracy of each information criterion in identifying the correct number of latent classes.

To evaluate performance, the simulation manipulates two central sources of model specification in mixture modeling—class enumeration and class proportion priors—across varying levels of sample size, class separation, and population-level class proportions. We examine how these factors interact to influence both convergence and class enumeration outcomes across all model selection indices.

7. Simulation Results

7.1. Convergence Rates

Convergence rates were calculated as the proportion of replications in which the highest potential scale reduction factor (across all estimated parameters within that replication) was below 1.05, indicating successful convergence, out of all replications considered across simulation conditions. While the mean convergence rate across all replications was 69.49% and the median convergence rate was 94.70%, convergence rates varied across simulation conditions, as presented in Figure 2.

In Figure 2, the columns represent six conditions derived from three sample size levels (150, 300, and 900) and two class proportion conditions at the population level (equal vs. unequal class proportions). The rows correspond to three levels of class separation based on the Mahalanobis distance

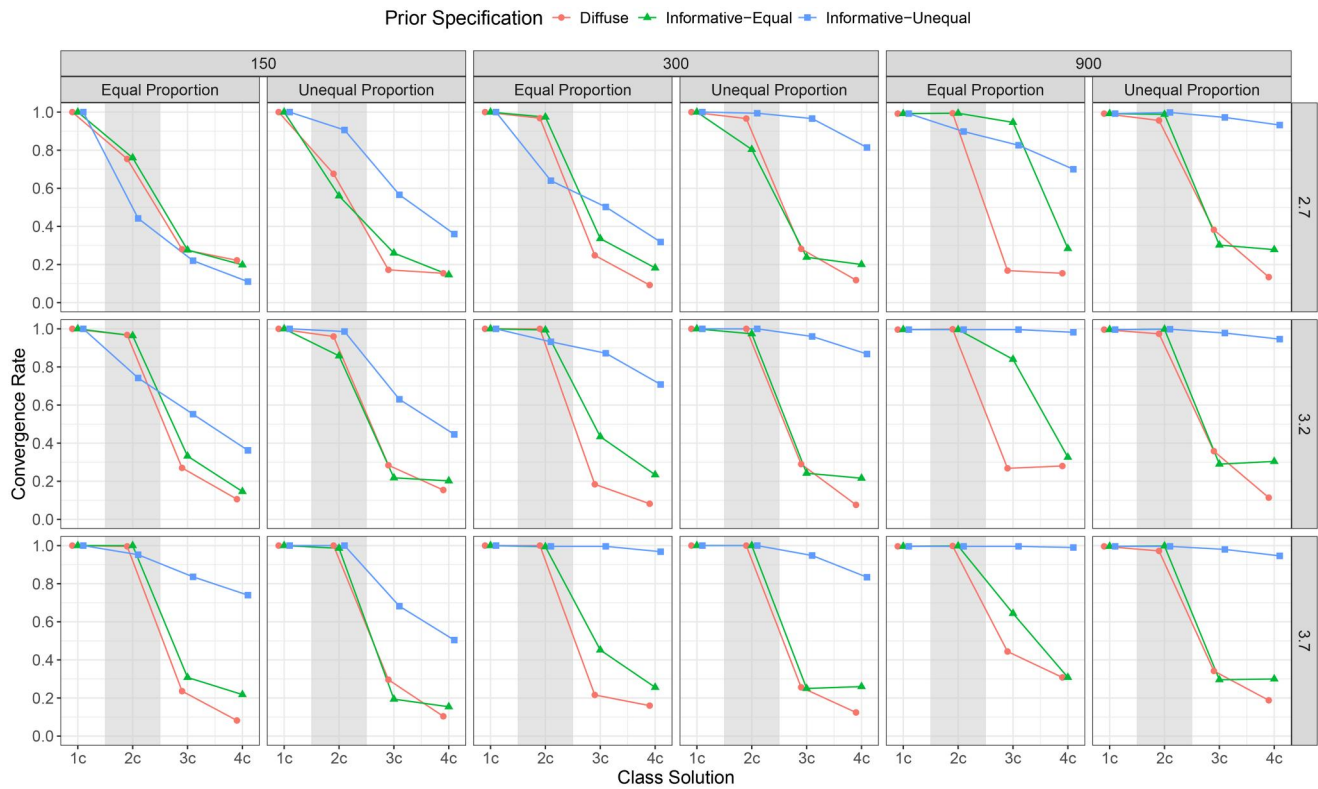


Figure 2. Convergence rates.

(MD = 2.7, 3.2, and 3.7). Within each plot, the x -axis represents the class solution, ranging from one to four classes, while the y -axis indicates the proportion of replications that successfully converged. Each plot contains three lines, distinguishing prior specifications: diffuse, informative-equal, and informative-unequal.

There are several general patterns in the convergence rates. First, a broad pattern emerged highlighting that convergence rates tended to drop as the number of estimated classes increased. For diffuse and informative-equal priors, this pattern was especially the case. We note, however, that for some informative-unequal prior conditions (especially when class proportions at the population level were unequal), convergence rates did not drop as much. We expect this to be due to the prior specification better aligning with the population model. The second pattern that emerged was that convergence rates increased as sample sizes got larger. Third, when prior specifications aligned with the class proportion at the population level, convergence rates were higher. Lastly, convergence rates generally increased as class separation increased; however, such an effect was particularly pronounced when class proportions were equal at the population level and in cases of overextraction (e.g., estimating three-class or four-class solutions).

For the upcoming results of selection rates, we included only replications in which all four class solutions (i.e., 1-through 4-class models) successfully converged. This ensured that model comparisons were conducted across a consistent set of competing solutions, rather than a partial set of converged solutions. This reflects typical research

practice in which comparisons are made only among class solutions that have reached convergence.¹

7.2. Selection Rates

Results for the simulation study are presented in Figures 3–5 for sample sizes $n = 150$, 300, and 900, respectively. The figures are all structured the same. Columns represent the different class separation conditions, with the smallest separation (MD = 2.7) on the left and the largest separation condition (MD = 3.7) on the right. There are six rows, with the top three rows aligning with the equal class proportion conditions and the bottom three rows aligning with the unequal class proportion conditions.

For the equal proportion conditions, row 1 aligns with diffuse priors, row 2 aligns with informative priors that assume equal class proportions (aligned with the true class proportion structure), and row 3 aligns with informative priors that assume unequal class proportions (misaligned with the true class proportion structure).

For the unequal proportion conditions, row 4 aligns with diffuse priors, row 5 aligns with informative priors that assume equal class proportions (misaligned with the true class proportion structure), and row 6 aligns with

¹We note that an alternative treatment of convergence, in which model selection indices were compared based on all available converged models within each replication (rather than requiring convergence across all four class solutions), was explored. Results under this approach are reported in the supplementary materials at the Open Science Framework repository (<https://osf.io/2resf/>).

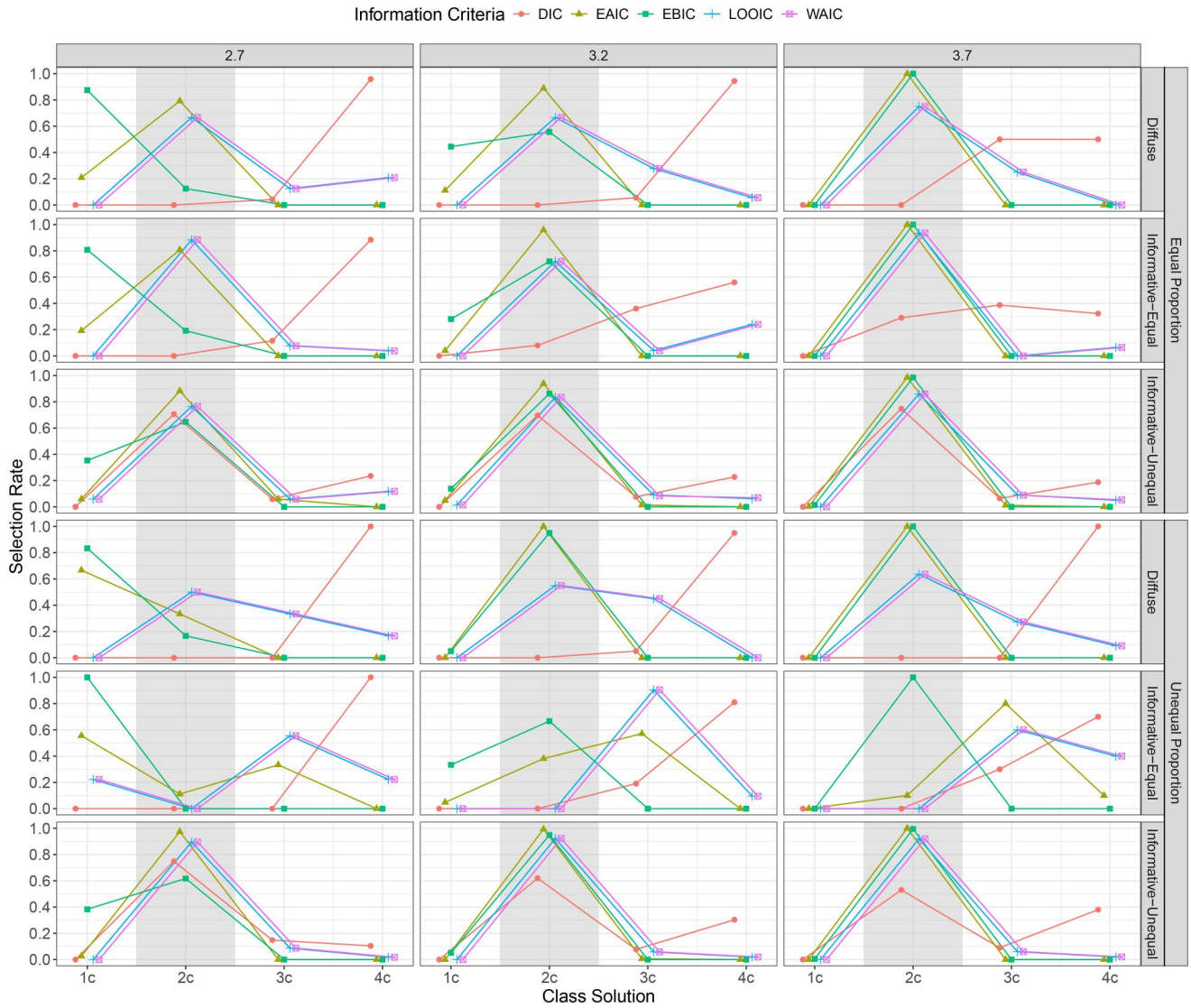


Figure 3. Selection rates for $n = 150$.

informative priors that assume unequal class proportions (aligned with the true class proportion structure).

Within each plot, there are five lines, each representing the model assessment indices under investigation. The outcome presented in these plots (the y -axis) represents the selection rates in terms of the proportion of replications selecting either a 1, 2, 3, or 4 class solution. There is a grayed shading in each plot highlighting that the true number of latent classes was 2 for all cells presented.

7.2.1. Equal Class Proportions, $n = 150$

In this section, we describe the results produced when equal class proportions were specified in the population model. For the 2-class structure, this implies an equal 50%/50% split of cases into the two latent classes at the population level. The first three rows of Figure 3 represent these cells for $n = 150$; sample size results for $n = 300$ and 900 are narrated in subsequent sections, but we focus initially on the smallest sample size, which presented the most nuanced results.

For the equal class proportion conditions, there are two prior conditions that align with the true structure of equal class proportions in the population. Row 1 represents diffuse priors, and row 2 represents informative priors reflecting equal class proportions. Although the diffuse prior settings are not at all informative, they do still make some assumption of equality in the specification of the latent class sizes. As a result, we have deemed these two prior conditions to be aligned with the true structure (with row 1 loosely aligned through the diffuse nature of the priors and row 2 more closely aligned with the informative nature of the priors).

Row 3 presents results from priors that are misaligned with the true class proportion structure. Specifically, these prior settings reflect a majority class with 70% of the cases and a minority class with 30% of the cases. The “aligned” versus “misaligned” prior specification types should be kept in mind when interpreting the results. In focusing on the first three rows, it is clear that there are differences in index performance. Likewise, there are some interesting interactions that result from class separation and prior type.

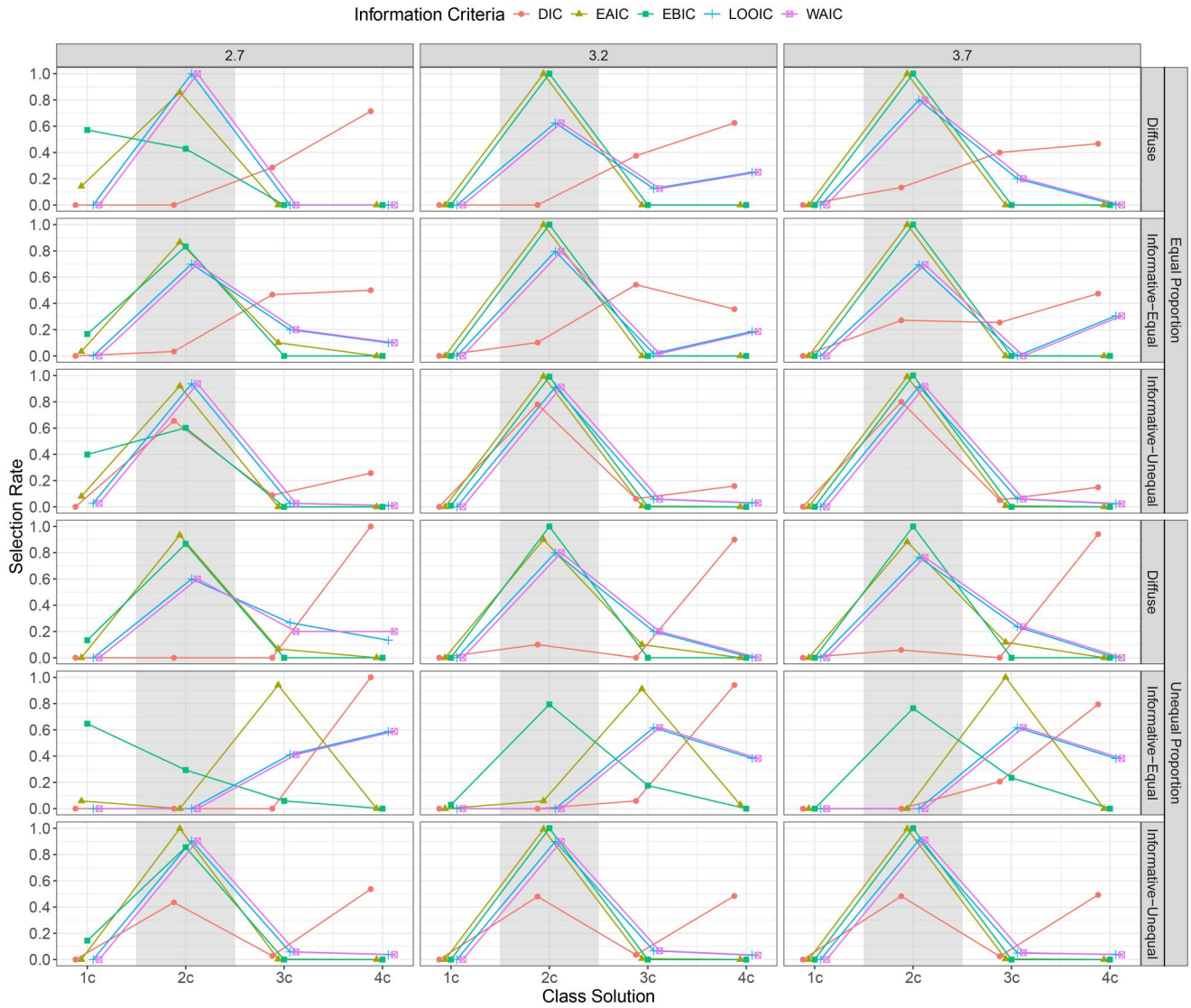


Figure 4. Selection rates for $n = 300$.

Overall, the EAIC, LOOIC, and WAIC showed consistent performance with just slight improvements as separation increased across the columns. The EBIC's performance was worse under the smallest separation condition, but it improved with moderate and high separation.

Perhaps the most interesting elements that were uncovered here were linked to the performance of the DIC. For the aligned priors in particular, the DIC yielded the worst overall performance with respect to selection rate recovery for the correct two-class solution. Although the ability to properly select two classes did improve as separation increased, the DIC still tended toward overextraction of the number of latent classes. That pattern was especially the case for the aligned prior types, namely for diffuse and informative-equal priors. When the prior was misaligned, and assuming an unequal class proportion, the DIC's performance improved. The informative priors assuming an equal class proportion across the classes produced similar results for the DIC as compared to the diffuse prior conditions. That was because both assumed equal class proportions. Albeit the diffuse settings were much less informative,

they still assumed an equality that aligned with the truth, which explained the similarity in results for these two prior settings.

7.2.2. Unequal Class Proportions, $n = 150$

The conditions with unequal class proportions in the population are presented in Rows 4–6 in Figure 3 for $n = 150$. Given the unequal proportions at the population level, two of the three prior conditions represented a misalignment with respect to the class proportions. Specifically, row 4 presents findings from diffuse priors which, although not informative, do contain a notion of equal class sizes. Row 5 presents results for the informative prior that assumed equal class proportions, which was a stronger misalignment with the true nature of the class structure. The last row in Figure 3 presents the informative prior with unequal class proportions, and it is aligned with the true structure of the class proportions at the population level (assuming a split of 70%/30%).

The EAIC and EBIC recovered the correct class solution more accurately as class separation increased. Patterns indicated that there was a preference for underextraction

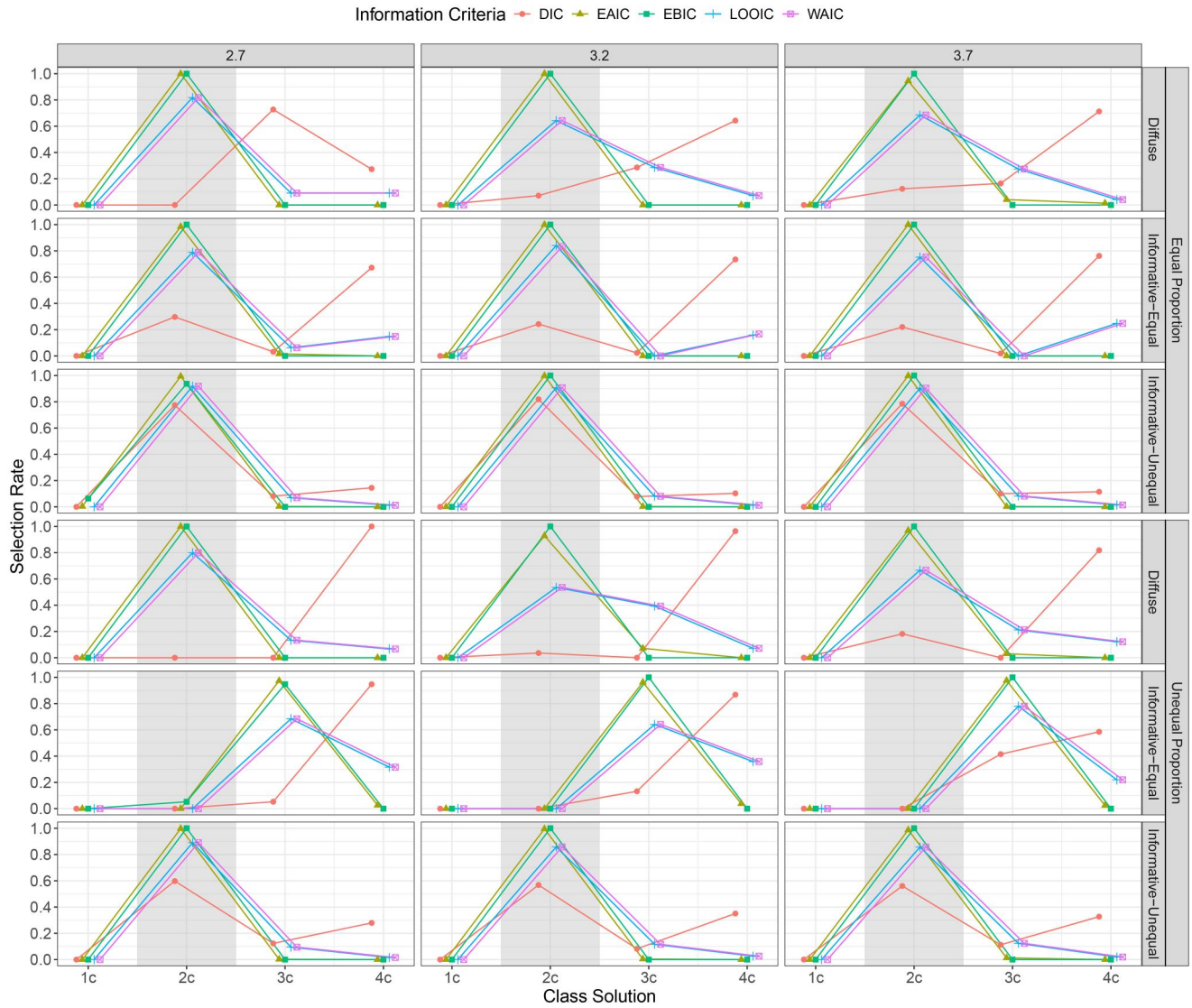


Figure 5. Selection rates for $n = 900$.

(1-class model) when class separation levels were poor, but the selection rate improved (favoring the 2-class solution) as class separation increased. Between the EAIC and EBIC, the EBIC was less sensitive to misaligned priors. The LOOIC and WAIC performed similarly to one another in these conditions of unequal class proportions. There was a higher tendency to overextract the number of classes with misaligned priors (diffuse and informative-equal settings).

Overall, the DIC was largely unaffected by class separation in these conditions, with some minor exceptions. Overextraction for the DIC was more likely under diffuse prior conditions, as well as the misaligned priors (i.e., informative with equal class proportions defined in the prior settings). In the case of the diffuse prior settings, the priors for the class proportions provided a weak equality assumption, and this assumption was enough to produce much higher levels of overextraction for the DIC. The DIC performed best under the prior setting that aligned with the true class structure (row 6), showing less tendency to overextract as compared to the misaligned priors in Rows 4 and 5. Note that although the DIC's selection rates were

improved under this prior condition, the accuracy rates were still lower compared to the other indices examined.

7.2.3. Comparison: Equal vs. Unequal Class Proportions, $n = 150$

One interesting pattern that emerged in the results was linked to the misaligned prior settings for the class proportions. For the equal class proportion conditions, the misaligned priors are in row 3 (informative and unequal proportions in the priors). For the unequal class proportion conditions, the misaligned priors are in row 4 (weak specification of equality in class sizes, albeit diffuse prior setting) and row 5 (informative priors with equal class proportions specified). As a comparison, we examined Rows 3 and 5 a bit closer since they each represented informative priors with incorrect class proportions specified. Row 3 pretty clearly shows that all indices were correctly aligned with the 2-class solution. Interestingly, row 5 shows, compared to row 3, a very different pattern of results. The findings indicated that incorrectly specifying equal class proportions as being unequal via the prior did not impact the ability of the

indices to select the correct model according to the population model. However, the impact of a misaligned prior was considerable when the population settings specified unequal latent classes, but the prior assumed equal classes. With the prior essentially ignoring that there was a true minority class and assuming equal sizes, the indices were much more likely to overextract the number of classes (with some instances of underextraction for the lowest class separation condition in the left plot of row 5).

7.2.4. Sample Size Considerations

The results presented above provided nuances for the smallest sample size of $n = 150$, which highlight the most variability in selection rate performance of the indices. However, we were also interested in examining the consistency of selection rate performance for each index across moderate and large sample sizes. The following sections detail the findings when sample sizes increased to a moderate size of $n = 300$ and to a large size of $n = 900$.

7.2.5. When $n = 300$ (Moderate Sample Size)

The patterns of index performance are presented in Figure 4. We first report the results for the equal class proportion condition, followed by the unequal class proportion condition. Within each condition, we examine whether the overall patterns changed compared to the smallest sample size condition ($n = 150$) and point out any notable differences. We also describe the results sequentially from the smallest to the largest class separation conditions. Finally, we report how these patterns varied across different prior conditions.

When the true class proportion was equal (first three rows of Figure 4) and the prior was diffuse, the overall pattern resembled what was observed at $n = 150$. However, the DIC performed better, with a reduced tendency to overextract the number of classes. More replications favored either a 2-class solution (in the case of the highest class separation at 3.7 MD) or a 3-class solution (across all class separation conditions). The EBIC did not perform well at the smallest class separation of 2.7 MD, but once class separation reached 3.2 MD, most replications favored the 2-class solution. The other indices (EAIC, LOOIC, and WAIC) consistently supported the 2-class solution across all levels of class separation. The results under the informative-equal prior condition largely mirrored those observed with the diffuse prior, with the exception that the EBIC could detect the true number of latent classes even at the smallest class separation (MD = 2.7). This pattern was consistent with the $n = 150$ condition, although, as with the diffuse prior, the DIC's tendency to overextract was mitigated. Finally, when the informative-unequal prior was used, all indices consistently supported the 2-class solution. These results were consistent with those observed at $n = 150$.

Next, we report the results for the case of unequal class proportions (last three rows of Figure 4). Under the diffuse prior, a noticeable difference when increasing the sample size from 150 to 300 was that, even at the smallest class

separation of 2.7 MD, most replications from all indices except the DIC correctly favored the 2-class solution. The DIC failed to select the 2-class solution and instead consistently overextracted the number of classes, with a preference for the 4-class solution across all class separation conditions. As such, the poor performance of the DIC in this condition remained the same as in $n = 150$, but for the other indices, performance improved compared to $n = 150$. When the informative-equal prior was used, which represents a large mismatch between the prior and the true unequal class proportion, only the EBIC tended to favor the true 2-class solution when the class separation was at least 3.2 MD. The other indices did not perform well in selecting the true solution. The poor performance of these indices was similar to that observed at $n = 150$. Under the informative-unequal prior specification, where there was a match between the true class proportion and prior settings, all indices except the DIC reliably favored the 2-class solution, regardless of class separation. The DIC did not show reliable performance, as it alternated between favoring the 2-class and the 4-class solutions. These patterns were again comparable to those observed in the smallest sample size condition.

7.2.6. When $n = 900$ (Large Sample Size)

Building on the previous comparison between cells with $n = 150$ and $n = 300$, we examined noticeable patterns that emerged as sample size increased from $n = 300$ (Figure 4) to $n = 900$ (Figure 5). First, we focus on cells with equal class proportions across both sample sizes, describing the impact of factors such as the degree of class separation and class proportion prior specification. Then, we move to cells with unequal class proportions.

In cells with equal class proportions (first three rows of Figure 5), conditions with misaligned (informative-unequal) priors yielded the best selection rates across indices regardless of sample size. Conversely, when using diffuse or correctly specified (informative-unequal) priors, the DIC tended to overestimate the number of latent classes. Although this tendency was noticeable at $n = 150$ and 300, it became more pronounced as the sample size increased. Moving on to other indices, the impact of class separation on the performance of the EAIC and EBIC differed across $n = 300$ and $n = 900$. At $n = 300$, the performance improved from a small to a moderate degree of class separation. In contrast, at $n = 900$, both indices consistently showed the highest selection rates for the correct class structure (i.e., 2-class), regardless of the class separation. In other words, given a large sample size ($n = 900$), class separation had nearly no impact on performance, compared to smaller sample sizes, which performed better under—at least—a moderate degree of class separation. Regarding the LOOIC and WAIC, diffuse and correctly specified (informative-equal) priors resulted in a higher rate of model overspecification across both sample size conditions compared to misaligned (informative-unequal) priors.

Shifting to cells with unequal class proportions (last three rows of Figure 5), the best performance across indices occurred in cells with informative-unequal priors—that is,

correctly specified class proportion priors. Specifically, the EAIC and EBIC performed best under diffuse and correctly specified (informative-unequal) prior settings. On the contrary, cells with misaligned (informative-equal) priors demonstrated the worst performance with a tendency to overextract the number of latent classes across indices. Notably, under misaligned (informative-equal) priors, the preference of all indices for over-specified solutions became more noticeable when moving from $n = 300$ to $n = 900$. This pattern was especially discernible for the EBIC. In the same vein, the selection rate of the DIC, LOOIC, and WAIC for over-specified solutions was greater under diffuse and mismatched (informative-equal) priors relative to correct priors (informative-unequal), suggesting a higher sensitivity of these indices to misaligned priors. Similar to the equal cells, the DIC tended to consistently overextract the number of latent classes.

Informative-unequal class proportion prior settings yielded the best performance across all indices—regardless of the actual class proportions (equal versus unequal)—for both sample sizes of ($n = 300$) and ($n = 900$). The DIC performed best under conditions of equal class proportions when using informative unequal class proportion priors. In summary, informative unequal prior settings provided the best overall performance across indices, irrespective of sample size or empirical class proportions, but the best results for the DIC were achieved with a combination of equal class proportions and informative unequal priors.

8. Secondary Investigation

As the primary simulation demonstrated a pronounced impact of class proportions on the performance of Bayesian model selection indices, it is important to examine whether these patterns generalize to even more extreme scenarios. To this end, we conducted a secondary simulation study incorporating a markedly imbalanced class structure (i.e., a 90/10 split), which reflects conditions that are plausible (e.g., Henson et al., 2007; Tueller & Lubke, 2010). This additional examination allows for a more practical evaluation of model selection behavior under severely unequal class proportions.

8.1. Simulation Design

We adopted the same data-generating model shown in Figure 1, using identical population parameter values as those listed in Table 1, except for the threshold values required to produce the 90/10 class proportion. The secondary simulation manipulated one extreme unequal class proportion condition (9:1 ratio), a single medium sample size (300), three levels of class separation ($MD = 2.7, 3.2$, and 3.7), and four candidate class solutions (1-class, 2-class, 3-class, and 4-class models). For the 2-, 3-, and 4-class models, three types of prior specifications were considered: diffuse, informative-equal, and informative-unequal. For the 1-class model, only a diffuse prior was used. These design factors yielded a total of 3 (from 1-class) + 27 (from

Table 3. Informative unequal prior specification for the 90/10 class proportion condition in the secondary simulation.

Class solution	Informative unequal
$n = 150$	
2-class	$\mathcal{D}(135, 15)$
3-class	$\mathcal{D}(135, 10, 5)$
4-class	$\mathcal{D}(135, 5, 5, 5)$
$n = 300$	
2-class	$\mathcal{D}(270, 30)$
3-class	$\mathcal{D}(270, 20, 10)$
4-class	$\mathcal{D}(270, 10, 10, 10)$
$n = 900$	
2-class	$\mathcal{D}(810, 90)$
3-class	$\mathcal{D}(810, 60, 30)$
4-class	$\mathcal{D}(810, 30, 30, 30)$

Note. The true number of classes was two, with 9:1 class proportions for extremely unequal conditions.

2–4-class) = 30 unique simulation conditions. The hyperparameter settings for the diffuse and informative-equal prior specifications were the same as those shown in Table 2, whereas the settings for the informative-unequal priors are summarized in Table 3. Each condition was replicated 100 times. While the number of replications was smaller than that in the primary simulation study, it was deemed sufficient for the purpose of evaluating whether patterns observed in the primary results generalize to more extreme conditions.

Data generation and Bayesian model estimation were conducted in *Mplus* Version 8.7 (L. K. Muthén & Muthén, 1998–2017). To avoid between-chain label switching, models were estimated using a single Markov chain with 40,000 iterations, with the first 20,000 discarded as burn-in. To prevent within-chain label switching, parameter identifiability constraints were imposed such that the intercept for Class 1 was constrained to be greater than that of Class 2 for models with two or more classes. To address potential issues with parameter estimates becoming trapped in local maxima, we supplied the same starting values for all class solutions as those used in the primary simulation study.

8.2. Simulation Results

8.2.1. Convergence Rates

Figure 6 presents convergence rates across class solutions for the secondary simulation with extremely unequal class proportions (90/10 split). Each replication was considered to have converged if the highest potential scale reduction factor across all parameters was less than 1.05. Overall, convergence was high for the 1-class solution under all prior specifications, but rates dropped considerably as the number of classes increased. For the 2-class models, convergence was somewhat stable with diffuse and informative-unequal priors, but markedly lower for the informative-equal priors, particularly at higher levels of class separation. For 3- and 4-class models, convergence rates declined sharply across all priors, with the lowest rates observed for the informative-equal specification. In contrast, the informative-unequal priors consistently yielded the most favorable convergence patterns, maintaining moderate rates even under the more complex class solutions.

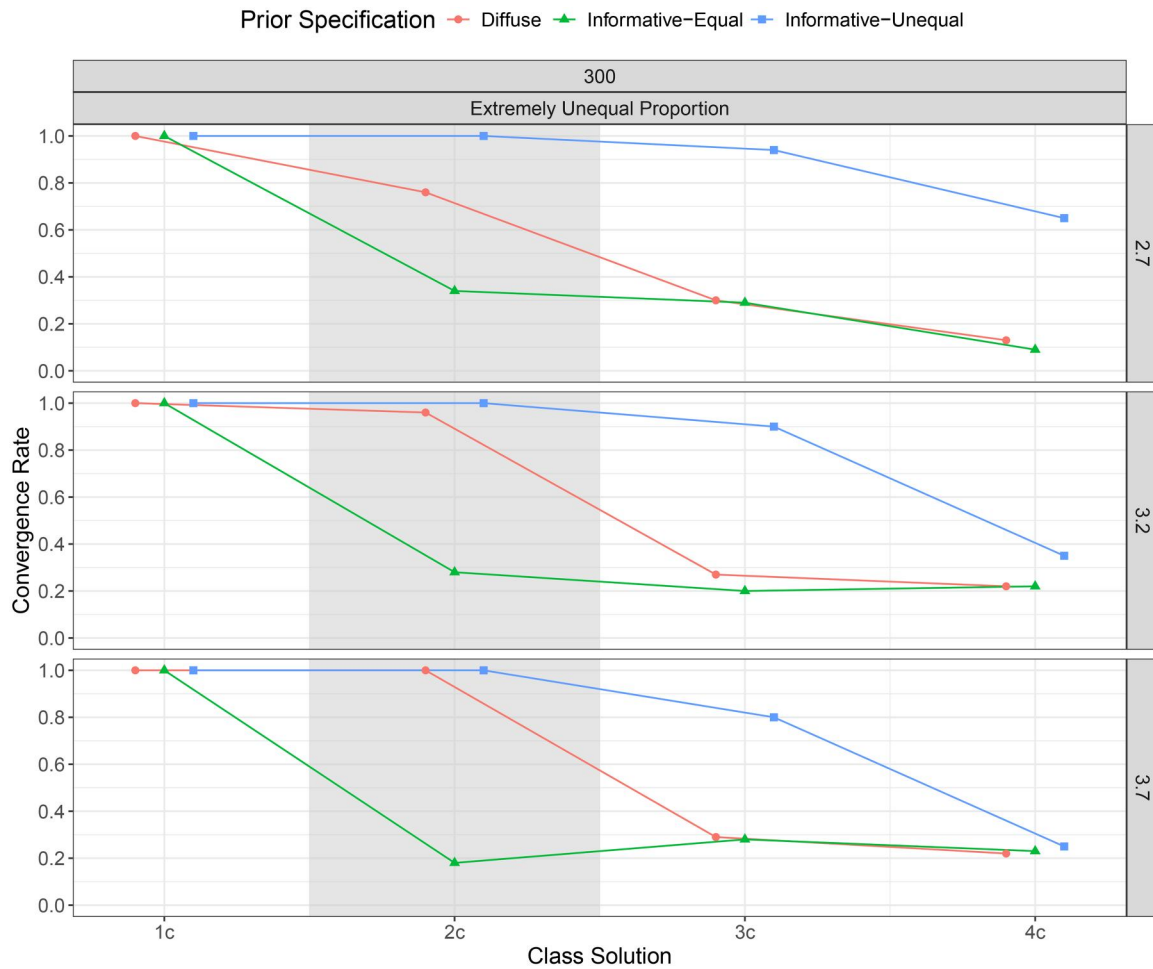


Figure 6. Convergence results for the secondary simulation.

These patterns suggest that the severe class imbalance created additional challenges for model estimation, especially when priors did not align with the underlying unequal class structure. However, the informative-unequal priors provided some protection against convergence failures relative to the diffuse and informative-equal priors.

8.2.2. Selection Rates

Figure 7 summarizes model selection rates under the extreme 90/10 class imbalance across levels of class separation and prior specifications. For the diffuse priors (top row), the information criteria were inconsistent, with selection spread across the 2-, 3-, and 4-class models. The EAIC and EBIC consistently selected the true 2-class model, and the LOOIC and WAIC experienced improvements as class separation increased (with the poorest separation condition tending to align with over-extraction for these indices). The DIC struggled under these conditions and had a strong tendency to over-extract.

For informative-equal priors (middle row), the true model solution is the least selected by all indices. The incorrect prior settings for the class proportions appear to strongly interact with the ability for indices to properly define the class structure.

Finally, for the informative-unequal prior settings (bottom row), we see very clear patterns of proper selection rates for all indices. The DIC still shows a slight tendency of over-extraction (especially clear for MD = 3.7), but the pattern of results is quite clear.

Overall, these findings highlight that class proportion priors play a critical role under severe imbalance: Informative-unequal priors consistently identified the true model, whereas diffuse and informative-equal priors led to unstable and often incorrect solutions.

9. Discussion

Several consistent patterns emerged across our findings. As anticipated, index performance improved substantially with increasing sample size, yielding clearer and more interpretable results in larger sample conditions (a pattern that aligns with statistical theory regarding power and precision).

The LOOIC and WAIC demonstrated superior performance in detecting class structure inconsistencies across most conditions. However, these indices struggled when faced with the specific challenge of unequal empirical class proportions paired with priors specifying equal class distributions. This misalignment between population characteristics and prior specifications proved particularly problematic for

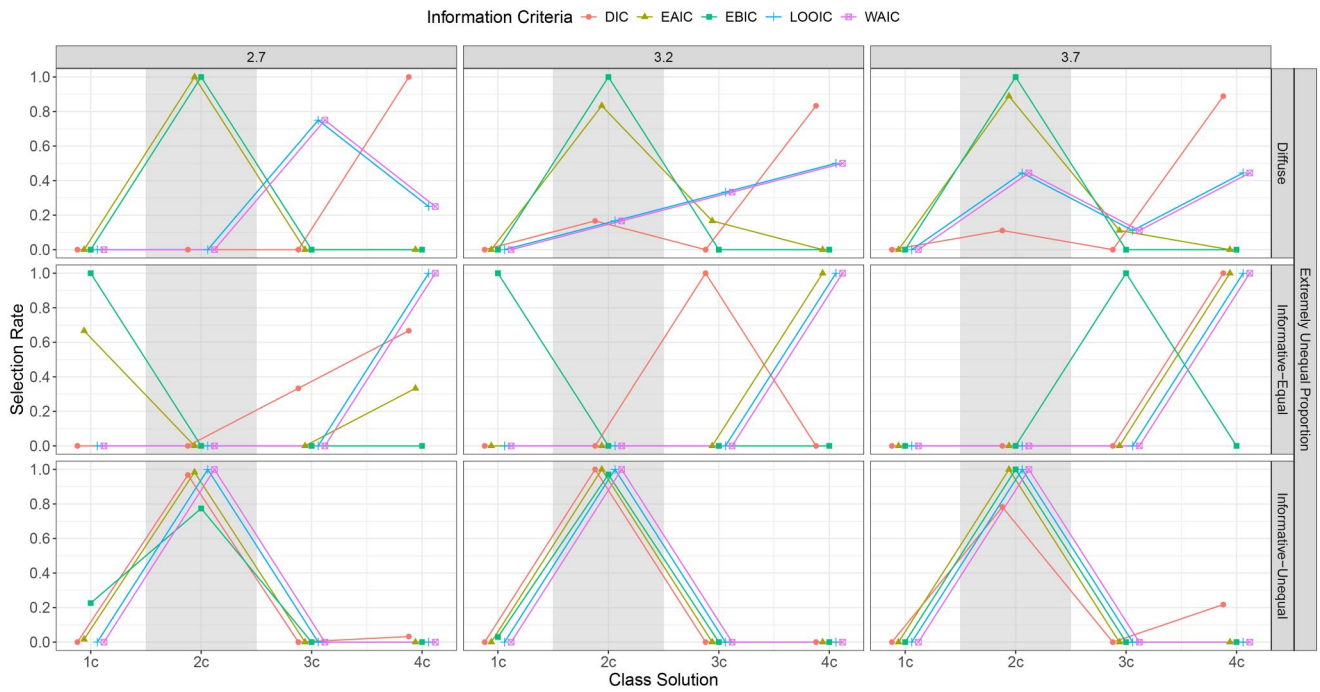


Figure 7. Selection results for the secondary simulation.

both the LOOIC and WAIC. Moreover, the impact of prior misalignment varied depending on whether the data contained a minority class or featured relatively equal class sizes. These findings highlight the critical importance of considering relative class proportions when selecting indices for determining appropriate class structure.

The EAIC emerged as a consistently reliable index across conditions, accurately detecting class structure inconsistencies in most modeling scenarios. In contrast, the EBIC demonstrated limitations in identifying correct class solutions, particularly when class separation was poor and sample size was small. Nevertheless, the EBIC performed adequately under conditions with more pronounced class separation, even when priors were misaligned in small and medium sample sizes. These patterns suggest that EBIC may still offer practical utility in specific contexts. This differential performance across separation conditions underscores the importance of considering multiple indices when making determinations about latent class structure, especially in datasets where class distinction may be ambiguous.

A particularly significant finding from our investigation concerns the performance of the DIC. Critically, this finding builds on earlier concerns about the shortcomings of the DIC (Spiegelhalter et al., 2014), providing new evidence of its limited robustness in the context of Bayesian mixture modeling. Our results demonstrate that the DIC exhibits a consistent tendency toward overextraction of latent classes. This overextraction rate increases with larger sample sizes under specific prior conditions, revealing an important three-way interaction effect between prior specifications, sample size, and index performance that researchers must consider to avoid latent class overextraction. Such overextraction can substantially alter substantive research conclusions by artificially splitting classes or incorrectly assigning

subjects to latent groups. Interestingly, there was one area where the DIC showed a clean solution in Figure 3. When the prior setting was misaligned in that it showed unequal class sizes but the true latent classes were equal in size, the DIC did not tend to overextract (row 3 of Figure 3). In this condition, the hyperparameter setting of the prior distribution down-weighted the possibility for overextraction. Specifically, high hyperparameter values in the Dirichlet prior exert a stronger influence, drawing more posterior mass toward the class with larger prior weights. These findings underscore the importance for applied researchers to exercise caution when interpreting DIC results and to potentially consider complementary indices when determining class enumeration in Bayesian mixture models.

To address concerns about “false” over-extraction (e.g., spurious classes of only 1–2%), we examined the estimated proportions of over-extracted classes in both the primary and secondary simulations. As summarized in Figure 8, the additional classes almost never appeared at such trivial sizes. In the 3-class models, over-extracted classes under diffuse priors were occasionally as small as ~10%, but more often were larger. Under informative-equal priors, spurious classes typically accounted for 20–40% of the sample. In the 4-class models, the fourth class was sometimes near 10% under diffuse priors but rarely below that level. Thus, the over-extracted classes we observed generally represented non-trivial proportions of the sample and were not easily dismissible as nuisance classes. Importantly, parameter estimates for these classes sometimes resembled existing classes but at other times differed, underscoring the need for researchers to incorporate substantive interpretability alongside fit indices when evaluating class solutions (see e.g., Nylund-Gibson & Choi, 2018).

Overall, the simulation findings from the primary and secondary investigations demonstrated that prior misalignment

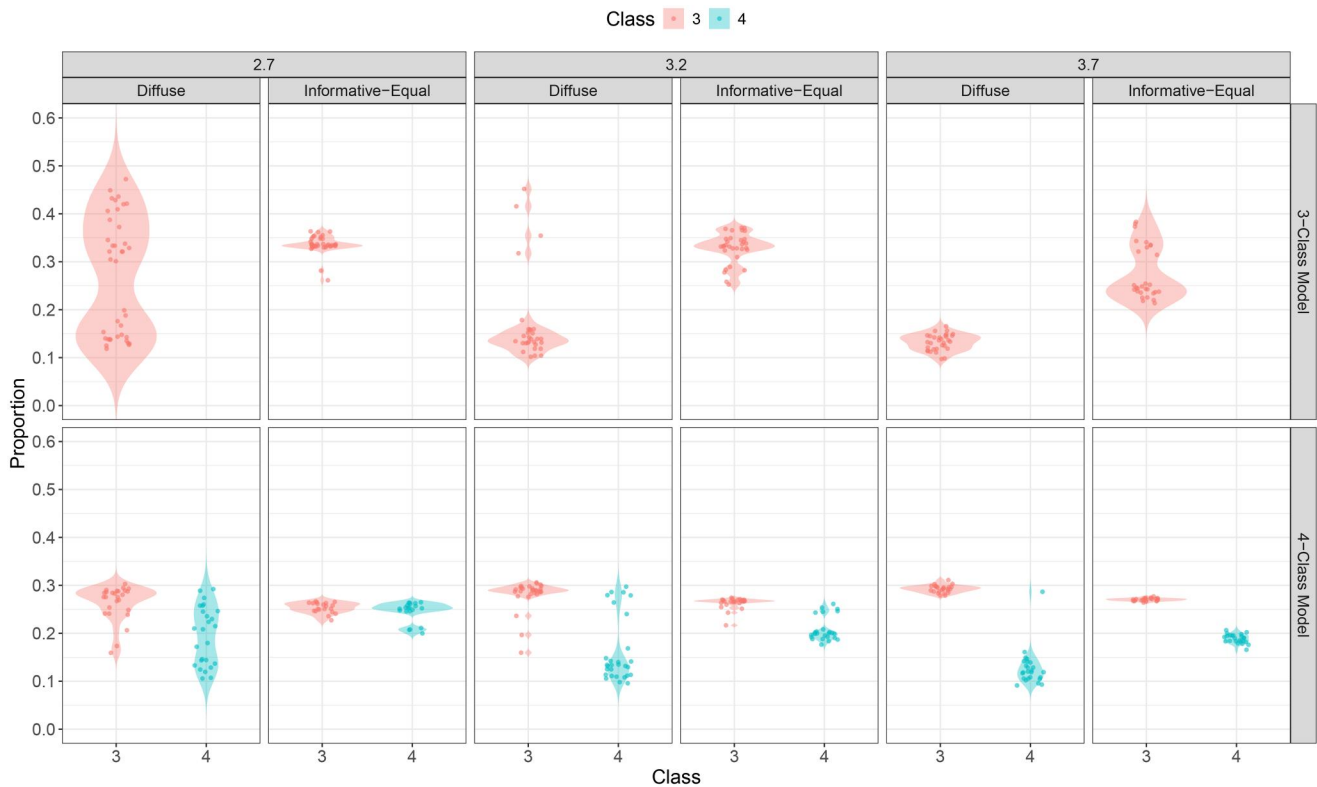


Figure 8. Proportion of overextracted classes in the secondary simulation.

did not simply reduce efficiency but can fundamentally distort the operating characteristics of the information criteria, making them appear unstable or misleading. For applied researchers, this implies that careful attention to prior specification is critical, particularly in settings where class imbalance is expected. In practice, incorporating prior knowledge about plausible class proportions (e.g., informed by theory or previous data) can substantially improve the reliability of information-criterion-based model selection decisions, whereas reliance on default or misaligned priors may lead to systematic misclassification of model complexity.

9.1. Future Research Directions to Consider

Class enumeration remains a critical methodological consideration for future investigations due to its profound impact on substantive interpretations and findings. An incorrect class solution inevitably leads researchers to draw conclusions and make generalizations that inadequately reflect the population(s) under study. While model selection tools can facilitate the selection of an appropriate class solution, our investigation (though revealing important performance patterns across indices) highlights the need for further research before we fully understand class enumeration and model selection within the Bayesian mixture modeling framework.

One promising direction for methodological advancement involves developing corrections or modifications to indices that demonstrated limitations in detecting the correct class solution. A potentially valuable approach would be adapting the complexity index in the DIC (specifically the p_D parameter) to address the overextraction issues.

Similarly, the EBIC, which proved particularly unreliable when class separation was poor, warrants reexamination. Its current formulation appears insufficiently sensitive to proper class structure, indicating a clear area for improvement.

Another valuable extension of this work would be comparing these indices' performance against alternative Bayesian modeling strategies. Specifically, semi- and non-parametric approaches for mixture models utilizing reversible jump Markov chain Monte Carlo or Dirichlet process algorithms have shown promising utility for latent mixture models (Ho & Hu, 2008; Ishwaran, 2000; Qiu et al., 2025; X.-Y. Song et al., 2011; X. Song et al., 2018; Yang & Dunson, 2010). However, these approaches tend to favor underextraction (X. Song et al., 2018). A comprehensive comparison would provide valuable insights regarding optimal analytical strategies: whether to employ model selection indices or estimate latent class numbers directly using specialized class estimation algorithms.

An additional area in need of future research is a more in depth investigation on how settings for the prior distributions for all model parameters (not just the class proportions) impact of the performance of the information criteria. Some previous work (see e.g., Depaoli, 2013; Lee & Harring, 2023) has indicated that priors placed elsewhere in the model (e.g., on growth parameter means, variances, or covariances) can impact parameter recovery. It follows that prior sensitivity analysis for these other model parameters may uncover performance patterns for the information criteria that impacts our understanding of model selection. We recommend that researchers examine this issue through a systematic sensitivity analysis of different prior forms and

hyperparameter settings on all model parameters. That work would create a more complete picture of index performance and the interaction between prior distributions and class enumeration.

A limitation of our current investigation is its exclusive focus on linear LGMMs. Future research should examine whether index performance patterns generalize across other SEM-based mixture models, including mixture SEM, latent class analysis models, nonlinear growth mixture models, and mixture confirmatory factor analysis (e.g., Heo et al., 2024; Tueller & Lubke, 2010; Whittaker & Miller, 2021; Yung, 1997). Additionally, our study did not address index performance in the presence of missing data, a significant omission given that research suggests indices can be substantially affected by missingness (see, e.g., Heo et al., 2024; Winter & Depaoli, 2022). This issue deserves thorough exploration in future work to develop a comprehensive understanding of these indices' performance in applied research contexts, where missing data represent the norm rather than the exception.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Sarah Depaoli  <http://orcid.org/0000-0002-1277-0462>
 Ihnwhi Heo  <http://orcid.org/0000-0002-6123-3639>
 Madelin Jauregui  <http://orcid.org/0009-0005-1689-359X>
 Haiyan Liu  <http://orcid.org/0000-0002-4272-9399>
 Fan Jia  <http://orcid.org/0000-0003-3855-532X>

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Bolfarine, H., & Bazan, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational and Behavioral Statistics*, 35, 693–713. <https://doi.org/10.3102/1076998610375834>
- Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An evaluation of PPP and DIC for Structural Equation Modeling. *Structural Equation Modeling*, 26, 39–50. <https://doi.org/10.1080/10705511.2018.1490648>
- Cao, C., & Liang, X. (2022). Sensitivity of fit measures to lack of measurement invariance in exploratory Structural Equation Modeling. *Structural Equation Modeling*, 29, 248–258. <https://doi.org/10.1080/10705511.2021.1975287>
- Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Chapman; Hall/CRC. <https://doi.org/10.1201/9781420057669>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Cassiday, K. R., Cho, Y., & Harring, J. R. (2021). A comparison of label switching algorithms in the context of growth mixture models. *Educational and Psychological Measurement*, 81, 668–697. <https://doi.org/10.1177/0013164420970614>
- Celeux, G., Forbes, F., Robert, C. P., & Titterton, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1, 651–673. <https://doi.org/10.1214/06-BA122>
- da Silva, M. A., Bazán, J. L., & Huggins-Manley, A. C. (2019). Sensitivity analysis and choosing between alternative polytomous IRT models using Bayesian model comparison criteria. *Communications in Statistics - Simulation and Computation*, 48, 601–620. <https://doi.org/10.1080/03610918.2017.1390126>
- Depaoli, S. (2021). *Bayesian structural equation modeling*. Guilford Press.
- Depaoli, S., Yang, Y., & Felt, J. (2017). Using Bayesian statistics to model uncertainty in mixture models: A sensitivity analysis of priors. *Structural Equation Modeling*, 24, 198–215. <https://doi.org/10.1080/10705511.2016.1250640>
- Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, 18, 186–219. <https://doi.org/10.1037/a0031609>
- Depaoli, S. (2014). The impact of inaccurate “informative” priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, 21, 239–252. <https://doi.org/10.1080/10705511.2014.882686>
- Depaoli, S., Jia, F., & Heo, I. (2023). Detecting model misspecification in Bayesian piecewise growth models. *Structural Equation Modeling*, 30, 574–591. <https://doi.org/10.1080/10705511.2022.2144865>
- Depaoli, S., Winter, S. D., & Liu, H. (2024). Under-fitting and over-fitting: The performance of Bayesian model selection and fit indices in SEM. *Structural Equation Modeling*, 31, 604–625. <https://doi.org/10.1080/10705511.2023.2280952>
- Du, H., Keller, B., Alacam, E., & Enders, C. (2024). Comparing DIC and WAIC for multilevel models with missing data. *Behavior Research Methods*, 56, 2731–2750. <https://doi.org/10.3758/s13428-023-02231-0>
- Fujimoto, K. A., & Falk, C. F. (2024). The accuracy of Bayesian model fit indices in selecting among multidimensional item response theory models. *Educational and Psychological Measurement*, 84, 217–244. <https://doi.org/10.1177/00131644231165520>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman; Hall/CRC.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016. <https://doi.org/10.1007/s11222-013-9416-2>
- Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling*, 14, 202–226. <https://doi.org/10.1080/10705510709336744>
- Heo, I., Depaoli, S., Jia, F., & Liu, H. (2024). Bayesian approach to piecewise growth mixture modeling: Issues and applications in school psychology. *Journal of School Psychology*, 107, 101366. <https://doi.org/10.1016/j.jsp.2024.101366>
- Heo, I., Jia, F., & Depaoli, S. (2024). Performance of model fit and selection indices for Bayesian piecewise growth modeling with missing data. *Structural Equation Modeling*, 31, 455–476. <https://doi.org/10.1080/10705511.2023.2264514>
- Ho, R. K. W., & Hu, I. (2008). Flexible modelling of random effects in linear mixed models—A Bayesian approach. *Computational Statistics & Data Analysis*, 52, 1347–1361. <https://doi.org/10.1016/j.csda.2007.09.005>
- Ishwaran, H. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, 371–390. <https://doi.org/10.1093/biomet/87.2.371>
- Kim, S., Tong, X., & Ke, Z. (2021). Exploring class enumeration in Bayesian growth mixture modeling based on conditional medians. *Frontiers in Education*, 6, 1–11. <https://doi.org/10.3389/educ.2021.624149>
- Kim, S., Tong, X., Zhou, J., & Boichuk, J. P. (2022). Conditional median-based Bayesian growth mixture modeling for nonnormal data. *Behavior Research Methods*, 54, 1291–1305. <https://doi.org/10.3758/s13428-021-01655-w>
- Kim, S.-Y. (2014). Determining the number of latent classes in single- and multiphase growth mixture models. *Structural Equation Modeling*, 21, 263–279. <https://doi.org/10.1080/10705511.2014.882690>

- Kim, S.-Y., Suh, Y., Kim, J.-S., Albanese, M., & Langer, M. M. (2013). Single and multiple ability estimation in the SEM framework: A non-informative Bayesian estimation approach. *Multivariate Behavioral Research*, 48, 563–591. <https://doi.org/10.1080/00273171.2013.802647>
- Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. L. (2015). Fitting a linear-linear piecewise growth mixture model with unknown knots: A comparison of two common approaches to inference. *Psychological Methods*, 20, 259–275. <https://doi.org/10.1037/met0000034>
- Lee, D. Y., & Harring, J. R. (2023). Handling missing data in growth mixture models. *Journal of Educational and Behavioral Statistics*, 48, 320–348. <https://doi.org/10.3102/10769986221149140>
- Liu, H., Depaoli, S., & Marvin, L. (2022). Understanding the deviance information criterion for SEM: Cautions in prior specification. *Structural Equation Modeling*, 29, 278–294. <https://doi.org/10.1080/10705511.2021.1994407>
- Liu, H., Heo, I., Depaoli, S., & Ivanov, A. (2025). Parameter recovery for misspecified latent mediation models in the Bayesian framework. *Structural Equation Modeling*, 32, 618–637. <https://doi.org/10.1080/10705511.2025.2475490>
- Liu, H., Heo, I., Ivanov, A., & Depaoli, S. (2025). Model assumption violations in Bayesian latent mediation analysis: An exploration of Bayesian SEM fit indices and PPP. *Structural Equation Modeling*, 32, 866–896. <https://doi.org/10.1080/10705511.2025.2503789>
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis* (1st ed.). Hall/CRC. <https://doi.org/10.1201/b13613>
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337. <https://doi.org/10.1023/A:1008929526011>
- Luo, Y., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling*, 59, 183–205.
- McNeish, D. (2023). A practical guide to selecting and blending approaches for clustered data: Clustered errors, multilevel models, and fixed-effect models. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000620>
- McNeish, D., & Harring, J. R. (2017). The effect of model misspecification on growth mixture model class enumeration. *Journal of Classification*, 34, 223–248. <https://doi.org/10.1007/s00357-017-9233-y>
- Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84, 802–829. <https://doi.org/10.1007/s11336-019-09679-0>
- Muthén, B., Khoo, S.-T., & Francis, D. (1998). *Multi-stage analysis of sequential developmental processes to study reading progress: New methodological developments using general growth mixture modeling*. CSE technical report 489. Retrieved April 14, 2025, from <https://eric.ed.gov/?id=ED427081>
- Muthén, B. (2001). Second-generation Structural Equation Modeling with a combination of categorical and continuous latent variables: New opportunities for latent class-latent growth modeling. In *New methods for the analysis of change* (pp. 291–322). American Psychological Association.
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Statistical analysis with latent variables: User's guide (version 8)*. American Psychological Association. <https://doi.org/10.1037/10409-010>
- Muthén, B. O., & Asparouhov, T. (2012). Bayesian Structural Equation Modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313–335. <https://doi.org/10.1037/a0026802>
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535–569. <https://doi.org/10.1080/10705510701575396>
- Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4, 440–461. <https://doi.org/10.1037/tps0000176>
- Peugh, J., & Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? a simulation study examining GMM's performance characteristics. *Structural Equation Modeling*, 19, 204–226. <https://doi.org/10.1080/10705511.2012.659618>
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) (pp. 1–10). <https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf>
- Qiu, M., Paganin, S., Ohn, I., & Lin, L. (2025). Bayesian nonparametric latent class analysis with different item types. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000728>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464. <https://doi.org/10.1214/aos/1176344136>
- Song, X., Kang, K., Ouyang, M., Jiang, X., & Cai, J. (2018). Bayesian analysis of semiparametric hidden Markov models with latent variables. *Structural Equation Modeling*, 25, 1–20. <https://doi.org/10.1080/10705511.2017.1364968>
- Song, X.-Y., Xia, Y.-M., Pan, J.-H., & Lee, S.-Y. (2011). Model comparison of Bayesian semiparametric and parametric Structural Equation Models. *Structural Equation Modeling*, 18, 55–72. <https://doi.org/10.1080/10705511.2011.532720>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76, 485–493. <https://doi.org/10.1111/rssb.12062>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64, 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Information Age Publishing.
- Tong, X., & Ke, Z. (2016). Growth curve modeling for nonnormal data: A two-stage robust approach versus a semiparametric Bayesian approach. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 229–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-38759-8_17
- Tong, X., & Ke, Z. (2021). Assessing the impact of precision parameter prior in Bayesian non-parametric growth curve modeling. *Frontiers in Psychology*, 12, 624588. <https://doi.org/10.3389/fpsyg.2021.624588>
- Tong, X., Kim, S., & Ke, Z. (2022). Impact of likelihoods on class enumeration in Bayesian growth mixture modeling. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology* (pp. 111–120). Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-04572-1_9
- Tong, X., & Zhang, Z. (2020). Robust Bayesian approaches in growth curve modeling: Using student's t distributions versus a semiparametric method. *Structural Equation Modeling*, 27, 544–560. <https://doi.org/10.1080/10705511.2019.1683014>
- Tueller, S., & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, 17, 165–192. <https://doi.org/10.1080/10705511003659318>
- van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistical modelling. *Nature Reviews Methods Primers*, 1, 1–26. <https://doi.org/10.1038/s43586-020-00001-2>
- van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian applications in psychology: The last 25 years. *Psychological Methods*, 22, 217–239. <https://doi.org/10.1037/met0000100>

- Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian Structural Equation Modeling. *Psychological Methods*, 23, 363–388. <https://doi.org/10.1037/met0000162>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27, 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594. <https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf>
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). Guilford Press.
- Whittaker, T. A., & Miller, J. (2021). Exploring the enumeration accuracy of cross-validation indices in latent class analysis. *Structural Equation Modeling*, 28, 376–390. <https://doi.org/10.1080/10705511.2014.915375>
- Winter, S. D., & Depaoli, S. (2022). Sensitivity of Bayesian model fit indices to the prior specification of latent growth models. *Structural Equation Modeling*, 29, 667–686. <https://doi.org/10.1080/10705511.2022.2032078>
- Yang, M., & Dunson, D. B. (2010). Bayesian semiparametric Structural Equation Models with latent variables. *Psychometrika*, 75, 675–693. <https://doi.org/10.1007/s11336-010-9174-4>
- Yung, Y.-F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62, 297–330. <https://doi.org/10.1007/BF02294554>