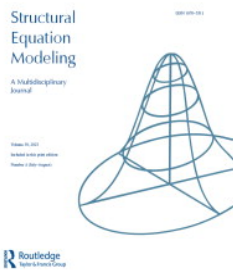

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hsem20

Detecting Model Misspecification in Bayesian Piecewise Growth Models

Sarah Depaoli, Fan Jia & Ihnwhi Heo

To cite this article: Sarah Depaoli, Fan Jia & Ihnwhi Heo (2023) Detecting Model Misspecification in Bayesian Piecewise Growth Models, Structural Equation Modeling: A Multidisciplinary Journal, 30:4, 574-591, DOI: 10.1080/10705511.2022.2144865

To link to this article: https://doi.org/10.1080/10705511.2022.2144865

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2023, VOL. 30, NO. 4, 574–591
https://doi.org/10.1080/10705511.2022.2144865

Routledge
Taylor & Francis Group

Detecting Model Misspecification in Bayesian Piecewise Growth Models

Sarah Depaoli, Fan Jia and Ihnwhi Heo

University of California

ABSTRACT

Bayesian estimation has become increasingly more popular with piecewise growth models because it can aid in accurately modeling nonlinear change over time. Recently, new Bayesian approximate fit indices (BRMSEA, BCFI, and BTLI) have been introduced as tools for detecting model (mis)fit. We compare these indices to the posterior predictive *p*-value (PPP), and also examine the Bayesian information criterion (BIC) and the deviance information criterion (DIC), to identify optimal methods for detecting model misspecification in piecewise growth models. Findings indicated that the Bayesian approximate fit indices are not as reliable as the PPP for detecting misspecification. However, these indices appear to be viable model selection tools rather than measures of fit. We conclude with recommendations regarding when researchers should be using each of the indices in practice.

KEYWORDS

Bayesian fit indices; model fit; model misspecification; model selection; piecewise growth models

Latent growth models (LGMs) represent a group of models that capture linear and nonlinear change over time. These models have gained in popularity as tools used to study developmental processes that are dynamic over time (see, e.g., Grimm et al., 2016). Many longitudinal processes are nonlinear by nature, thus giving rise to increased demand in models that can estimate dynamic change. The piecewise LGM is one example that can handle nonlinear growth through modeling distinct growth phases, which are joined together by *knots*. Piecewise LGMs are commonly implemented in longitudinal studies that examine nonlinear processes over time (see, e.g., Chung et al., 2017; Lee & Rojewski, 2009; Jaggars & Xu, 2016; Patrick & Schulenberg, 2011).

One pitfall of LGMs emerging from the methodological literature is that nonlinear trends can be difficult to properly estimate (Diallo et al., 2014), and inaccurate estimates of the growth patterns can be obtained. However, the use of Bayesian methods for estimation has proven to be an advantageous approach over the conventional frequentist framework (Depaoli, 2013; Smid, McNeish, et al., 2020). The Bayesian approach combines information via a prior distribution with the sample data to form a resulting posterior distribution. For example, using a simple linear growth model, information can be derived from previous research to form priors for the latent intercept mean and the latent linear slope mean. The information captured through the prior distribution is a key ingredient that can help improve the accuracy of the final model estimates produced, thus improving the utility of LGMs when used in substantive settings. Typically, the parameters of most interest are the means and variances for the latent growth factors (e.g., the latent intercept and slope terms), as well as the latent factor covariance matrix. Overall, these latent factors are important to accurately estimate because they produce the growth trajectory representing the estimated growth or change patterns over time.

Incorporating prior information can help to improve the accuracy of LGM estimates, even when samples are relatively smaller in size (McNeish, 2016) and when growth is nonlinear in nature (Depaoli & Boyajian, 2014; Kohli & Harring, 2013; Winter & Depaoli, 2022). Given that relatively smaller samples and nonlinearity are two common issues within longitudinal research implementing LGMs, it is important to fully explore the performance of Bayesian methods for these models and highlight the potential value of Bayesian estimation over traditional frequentist approaches.

One element within the Bayesian implementation of LGMs that needs further examination is the use of model fit and comparison indices. Within LGM research, it is common for researchers to use model fit or comparison measures to assess multiple competing models to select the one that "best" represents the patterns captured by the data (Chou et al., 1998; Kroese et al., 2013; Li et al., 2019; Wu et al., 2009). Until recently, the model selection and fit index choices within the Bayesian framework were quite limited, largely focusing on the posterior predictive *p*-value (PPP-value; Gelman et al., 1996). However, there has been a new expansion of Bayesian model fit tools available with the extension of conventional approximate model fit indices into the Bayesian framework (Asparouhov & Muthén, 2021; Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018).

Thus far, methodological research has indicated that these indices have great potential to aid in the detection of model misspecification for structural equation models

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

Supplemental data for this article is available online at https://doi.org/10.1080/10705511.2022.2144865.

(Asparouhov & Muthén, 2021; Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018). However, the literature is sparse regarding their performance for LGMs. To our knowledge, one paper has examined the performance of these indices in the context of LGMs. Specifically, Winter and Depaoli (2022) found (for a quadratic model) that priors that diverged from the population values impacted the performance of Bayesian model fit and selection tools such that correctly specified LGMs appeared misspecified. This was an important finding in that it highlighted the influence that priors can have on the performance of these approximate fit indices. However, there is a complete lack of methodological work examining the performance of these indices in the context of piecewise LGMs. Our goal is to extend the work conducted on LGMs to explore the ability of these indices to detect model misspecification in piecewise trajectories and the impact of priors. This goal is rooted in notable gaps existing in the methodological literature. Specifically, it has been established that prior specifications have an important role in properly detecting and capturing degrees of nonlinearity (e.g., Lock et al., 2018). Prior specification can also influence the performance of Bayesian model fit measures in various longitudinal models (Cain & Zhang, 2019; Winter & Depaoli, 2022). Therefore, it is interesting to examine the potential impact of priors and the ability of these Bayesian fit measures to detect model misspecification in the presence of piecewise trajectories.

1. Goals and Organization of the Current Investigation

The current study examines these issues in the context of the piecewise LGM, which can be used to capture nonlinearity over time. Specifically, we present a simulation study that examines the overall performance of the Bayesian model fit measures in terms of model (mis)specification and prior specification. Our goal is to uncover performance and provide recommendations for applied researchers who are looking to use the Bayesian framework to assess nonlinear growth via piecewise models, which is likely to involve assessing model fit. We will compare the performance of the PPP-value to the new Bayesian approximate fit indices, as well as two common model comparison indices: the Bayesian information criterion (BIC; Schwarz, 1978) and the deviance information criterion (DIC; Spiegelhalter et al., 2002). This investigation is essential to uncover the accuracy and ability for these indices to properly detect model misspecification for piecewise LGMs. This study will help to answer the following two questions regarding model and prior specification:

1. (Model Misspecification) How well do Bayesian model fit and asssessment measures detect model misspecification for piecewise LGMs?
2. (Prior Specification) Does prior specification (e.g., informativeness and accuracy of the prior) have an impact on the overall performance of model fit indices for piecewise LGMs?

A variety of conditions are examined here to fully expose specification issues tied to the model and the priors. The current paper is organized as follows. First, we present the basic form of the LGM and then extend this to the piecewise LGM. We discuss issues tied to knot placement and the general state of knowledge surrounding nonlinear modeling. This is followed by details surrounding the Bayesian implementation of the piecewise LGM, including a presentation of the model fit and comparison indices explored here. We then include a discussion on the performance of model fit and comparison indices in the LGM literature. Next, we present the simulation design, which is followed by a presentation of the results. The paper concludes with a discussion of when Bayesian fit and model comparison indices can (and cannot) be trusted to detect model misspecifications. We present recommendations of use for applied researchers, as well as future methodological directions that are needed to improve the assessment of fit for Bayesian (piecewise) LGMs.

2. Latent Growth Models: Basics and the Piecewise Extension

LGMs estimate overall growth trajectories through repeated measures of a group of individuals, and simultaneously allow for individual variability in the trajectories. In these models, group growth trajectories are captured by the means of growth factors (i.e., latent intercept and latent slope), while the individual variabilities are measured by the variances of these growth factors. Repeated measures are treated as multiple indicators of the latent factors. The factor loadings vary depending on the shapes of trajectories. An LGM can be represented in matrix notation:

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\boldsymbol{y}_i$ is a vector of T repeated measures for individual i, $\boldsymbol{\eta}_i$ is a vector of m latent factors for individual i, $\boldsymbol{\Lambda}$ is a $T \times m$ matrix of the factor loadings, and $\boldsymbol{\varepsilon}_i$ is a vector of T residuals that cannot be explained by the trajectory. For example, in a simple linear latent growth curve model, two latent factors ($m = 2$) represent a latent intercept (η_{0i}) and a latent linear slope (η_{1i}) for individual i. In this case, Equation 1 can be rewritten as follows:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Ti} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_1 \\ 1 & \lambda_2 \\ \vdots & \vdots \\ 1 & \lambda_T \end{bmatrix} \begin{bmatrix} \eta_{0i} \\ \eta_{1i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \vdots \\ \varepsilon_{Ti} \end{bmatrix}. \tag{2}$$

Equation 2 also has an expanded form that represents y_{ti}, the observed variable for individual i at time t, as a function of two latent factors (η_{0i} and η_{1i}) and the residual ε_{ti}:

$$y_{ti} = \eta_{0i} + \eta_{1i}\lambda_t + \varepsilon_{ti}. \tag{3}$$

The factor loadings for the latent intercept are all fixed at 1 because the intercept does not change over time; the loadings associated with the linear slope are usually fixed to a linear progression of time scores that associates with the repeated measures (λ_t, $t = 1, 2, \ldots, T$). The linear slope loading for the first occasion is typically set at 0, therefore, we can have $\lambda_t = (0, 1, 2, 3, 4)'$, for five equally spaced measurement occasions ($T = 5$). If the time score is centered at the midpoint, then $\lambda_t = (-1, -2, 0, 1, 2)'$ (Grimm et al., 2016).

Linear LGMs can be extended to estimate nonlinear trajectories that include multiple growth phases. For example, learning process often demonstrates a rapid growth in the beginning and then a slower growth in the later phase; and the shape of the growth trajectory may change due to an intervention (Kohli et al., 2015). Piecewise LGMs allow researchers to model change processes with distinct phases (Kohli & Harring, 2013). In a piecewise LGM, two adjacent growth phases are separated by a change point or knot, which can be known (fixed) or freely estimated. Considering a two-phase linear-liner piecewise process, the two phases of growth are captured by two linear slope growth factors. Suppose the knot is known to be at the k^{th} occasion, then we can adjust Equation 3 to represent the piecewise growth model:

$$y_{ti} = \eta_{0i} + \eta_{1i} \times \min(\lambda_t, \lambda_k) + \eta_{2i} \times \max(\lambda_t - \lambda_k, 0) + \varepsilon_{ti}, \tag{4}$$

Where η_{0i}, η_{1i}, and η_{2i} are latent intercept, and the first and second latent linear slopes, respectively; λ_t is the time score associated with the occasion t ($t = 1, 2, \ldots, T$), and λ_k is the time score associated with the knot at the k^{th} occasion. For occasions where $\lambda_t \leq \lambda_k$, the first and second latent slopes have loadings of λ_t and 0, respectively; for $\lambda_t > \lambda_k$, their loadings becomes λ_k and (λ_t - λ_k). For example, when there are seven measurement occasions and the knot is at the 4th occasion, then $k = 4$, $\lambda_k = 3$, and the $\mathbf{\Lambda}$ matrix is written as follows:

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 3 & 1 \\ 1 & 3 & 2 \\ 1 & 3 & 3 \end{bmatrix}. \tag{5}$$

An example of a path diagram of the piecewise LGM is pictured in Figure 1 and is the focus of the current investigation. In addition, we have included examples for what piecewise growth trajectories look like in Figure 2; this figure will be further explained in the Simulation Design section.

The piecewise LGM can also accommodate a variation of more complex growth scenarios. One example is the piecewise LGMs with disjointed knots (Cudeck & Codd, 2012; Rioux et al., 2021). In addition, Harring et al. (2021) discussed three variants of piecewise LGMs with both fixed and freely-estimated knot(s), including the three-phase linear model, segmented polynomial model, and piecewise model with exponential functions. These advanced piecewise LGMs are beyond the scope of the current investigation, but we mention them to provide the context of how versatile the piecewise LGM can be.

2.1. Model Misspecification and the Piecewise LGM

Relevant to the current investigation is the (mis)specification of the piecewise LGM. As with any latent variable model, assessing model fit and the ability to properly detect specification errors is a key component to proper specification and model interpretation. There are many ways in which the piecewise LGM can be misspecified. Perhaps one of the most obvious specification errors is if the nonlinearity of the trajectory was misspecified to be linear by ignoring the presence of the knot. In this instance, the piecewise model would be misspecified to a linear LGM. Related to this situation, Leite & Stapleton (2011) studied growth trajectory misspecification, in part, by fitting a linear growth model (analysis model) to data generated from a piecewise model (population model). They noted that misspecifying the growth trajectory in this way (e.g., by misspecifying a nonlinear trajectory to be linear) is akin to simultaneously misspecifying the mean and covariance structure of the model. The consequences are that incorrect trajectory estimates can be obtained, potentially leading toward substantively different conclusions than what exist in the population.

Another form of misspecification in piecewise LGMs is to incorrectly specify the knot location in the trajectory. The model itself would still produce a piecewise trajectory, but it would be incorrect in its precise nonlinear formation. In practice, the exact location of the knot may be unknown. For example, if an intervention is administered at time-point 4, that may not be the turning point in the trajectory because there could be a delayed reaction to the intervention (in this case, perhaps time-point 5 would be the turning point where the knot should be placed). Ning & Luo (2017) discussed this issue, indicating that misspecifying the knot location may lead to incorrect conclusions surrounding the growth rates.

The current investigation extends these concepts of model misspecification into the Bayesian framework, assessing the influence that priors may have in this modeling process. In addition, we examine the ability of Bayesian indices to detect this type of misfit, which may have important substantive consequences.

2.2. Relevant Prior Distributions

For the Bayesian implementation of the piecewise LGM, there are several priors that are of importance. First, latent growth parameters (η_0, η_1, and η_2) have estimated means that are assumed to be distributed normally as $\mathcal{N}(\mu, \sigma^2)$, where μ represents the mean hyperparameter and σ^2 represents the variance hyperparameter. These hyperparameters control the location and level of informativeness (or precision), respectively. The latent growth parameters are typically allowed to covary, and the prior for this latent factor covariance matrix (with growth factor variances on the diagonal and covariances on the off-diagonal) can be defined through the inverse Wishart distribution as $\mathcal{IW}(\mathbf{\Psi}, \nu)$. The $\mathbf{\Psi}$ hyperparameter is a positive definite matrix, which can

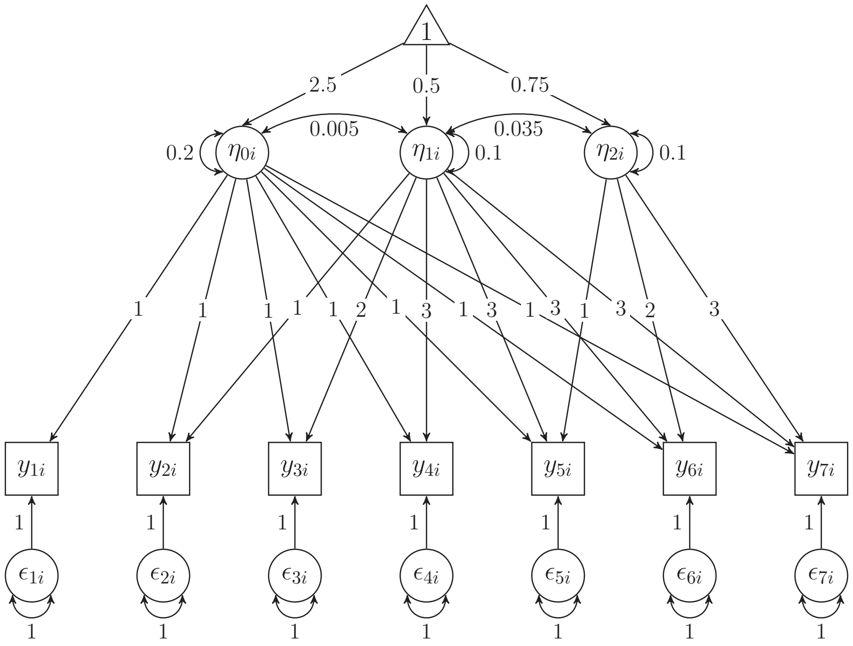

Figure 1. Piecewise growth curve model for the largest slope change condition.

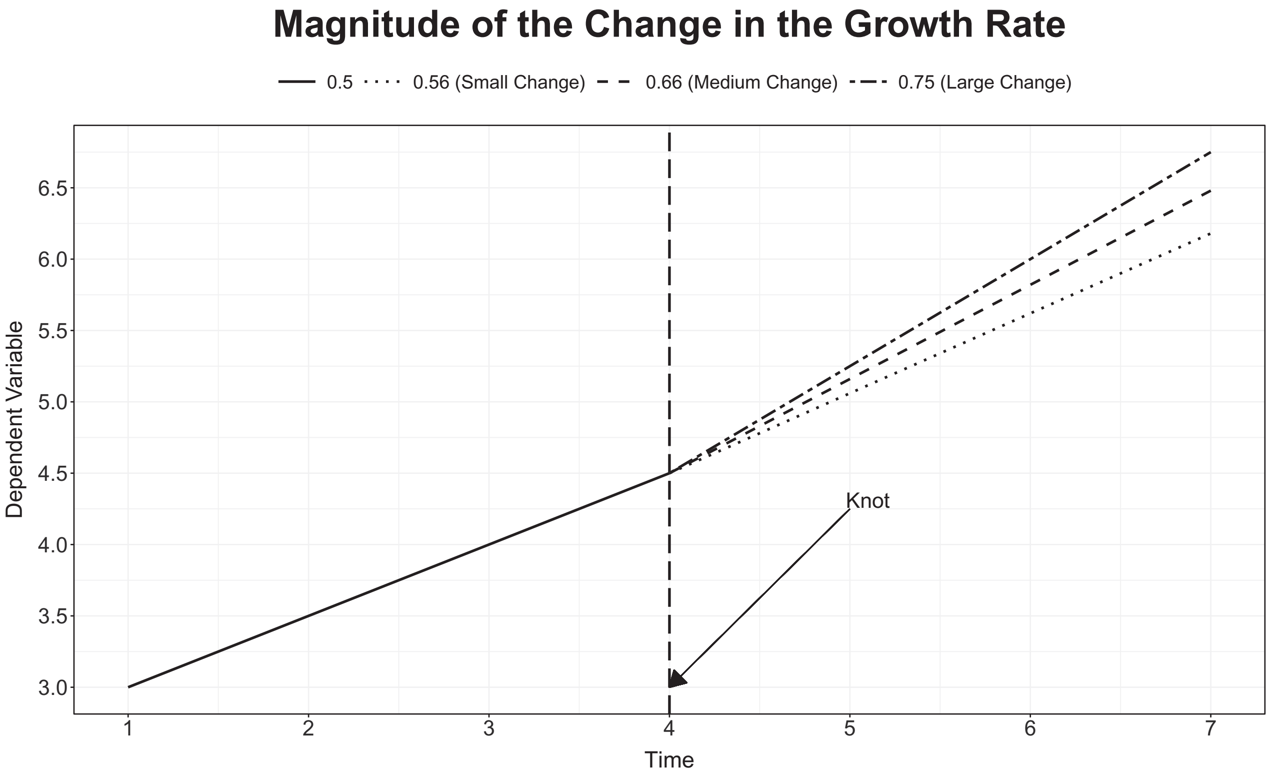

Figure 2. Three conditions of growth trajectories.

be defined in a variety of ways, including as an identity matrix. The ν hyperparameter represents the degrees of freedom, and the specification of $\mathbf{\Psi}$ and ν control the level of informativeness of this prior for the latent factor covariance matrix. The last prior that is typically specified for this model is for the variances in the model, which include the error variances that are tied to the repeated measures data. This prior can be specified in a variety of ways, but the conventional distributional form for a variance is to use an inverse gamma prior such as $\mathcal{IG}(a, b)$. In this case, the hyperparameters a and b are shape and scale parameters for the inverse gamma distribution, respectively. As detailed in the Design section below, the only prior that is manipulated in the current investigation is the prior placed on the growth factor means. The other priors were left as default settings implemented in the M*plus* software (Muthén & Muthén, 1998–2017).[1]

3. Model Fit and Selection Indices

3.1. Information Criteria

Although there are many information criteria to choose from, our investigation will examine two in detail: the BIC and the DIC. These two criteria are commonly implemented within the Bayesian SEM context for model selection purposes. Nested or non-nested models can be compared with each of these indices.

The BIC is based on an approximate of the Bayes factor and is written as follows:

$$BIC = (-2)\log(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) + q\log n, \tag{6}$$

where the number of model parameters (represented by $\boldsymbol{\theta}$) is q, the sample size is n, and $\tilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ depending on $\boldsymbol{y}$.

The second index we will cover here is the DIC, which includes the *effective number of parameters* (which is often lower than the actual number of model parameters) when determining model complexity, as specified by the user (Spiegelhalter et al., 2002); the authors termed this as choosing the "focus." In a Bayesian analysis, the number of parameters counted comprising the idea of model complexity can increase rapidly as more and more hyperparameters (or hyperpriors placed on hyperparameters) are included in the model; each model prior is comprised of hyperparameters, adding to the number of parameters in the model. The DIC allows the researcher to remove hyperparameters from the count, which removes them from the measure of model complexity, and makes the DIC a better index for handling a larger number of hyperparameters.

Before defining the DIC, we must first define the *deviance*, seen as:

$$D(\boldsymbol{\theta}) = -2\log(f(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log(h(\boldsymbol{y})), \tag{7}$$

where $h(\boldsymbol{y})$ is a standardized term which is a function of the data $\boldsymbol{y}$. Next, the *effective number of parameters*, p_D, is written as follows:

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\tilde{\boldsymbol{\theta}}), \tag{8}$$

where $\tilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ depending on data $\boldsymbol{y}$, and $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance, which can be defined as follows:

$$\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta}}\left[-2\log(f(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{y})\right] + 2\log(h(\boldsymbol{y})). \tag{9}$$

These measures form the DIC for model comparison as follows:

$$\begin{aligned} DIC &= \overline{D(\boldsymbol{\theta})} + p_D \\ &= D(\tilde{\boldsymbol{\theta}}) + 2p_D \\ &= 2\overline{D(\boldsymbol{\theta})} - D(\tilde{\boldsymbol{\theta}}). \end{aligned} \tag{10}$$

As long as $D(\boldsymbol{\theta})$ can be computed in closed form (e.g., there are no missing data present, which is the assumption we use here), then $\overline{D(\boldsymbol{\theta})}$ can be approximated using Markov chain Monte Carlo (MCMC) and taking the mean of the simulated values of $D(\boldsymbol{\theta})$.[2]

3.2. Posterior Predictive p-Value (PPP-Value)

The posterior predictive model check (PPMC) process is a popular procedure used to assess model fit within the Bayesian estimation framework. Before delving into the technical details, we will provide a brief conceptual overview of the three steps used in the PPMC process. First, an observed data posterior distribution is used to derive parameter estimates (e.g., based on the mean of the posterior). Second, each MCMC iteration (i.e., each sample in the Markov chain) generates a replicated dataset the same size as the observed dataset. Finally, a discrepancy function is computed between the observed and replicated data. A reference distribution is used to evaluate the extremeness of the observed data test statistic. Model misspecification is identified through a low PPP-value, indicating notable discrepancy between the observed and replicated data.

As noted above, the PPMC process allows a researcher to examine how consistent the observed data are with the proposed model. Specifically, it assesses whether *expected* data from the model are consistent with the *observed* sample data (Stern & Cressie, 2000). For a well-fitting model, the discrepancy between the fit of the model to the observed data and to the replicated data should be minor. However, as model misfit increases in severity, this discrepancy will also increase.

Examining this discrepancy first starts with taking draws from the posterior predictive distribution tied to the replicated data $\boldsymbol{y}^{rep}$, and this can be written as follows:

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{11}$$

[1] For the interested reader, there are additional methods for specifying prior distributions on the parameters included in this investigation. For example, separation strategy priors can be used for the latent factor covariance matrix (Depaoli, 2021; Liu et al., 2016), and there are a variety of prior settings that can be used for variance parameters (Gelman, 2006). To keep the simulation conditions manageable, and focus on the priors most commonly modified in applied research, we opted to only examine different settings for the latent growth factor means.

[2] It is important to note that sometimes the DIC is considered to only be *partially* Bayesian because it does not use the entire posterior. Rather, the DIC uses the mean of the simulated values from $D(\boldsymbol{\theta})$. As a result, many other indices have been developed within the Bayesian estimation framework. However, we restrict our investigation to these indices because they are currently the most widely used. For more information on the shortcomings of the DIC, see Spiegelhalter et al. (2014). In addition, the DIC is computed differently in popular Bayesian software packages (Merkle et al., 2019). For example, M*plus* and the R blavaan package compute the marginal DIC, in which the likelihood component is integrated over the latent variables. Other software packages, such as BUGS and JAGS, use the conditional DIC, in which the likelihood is conditional on the latent variables. The magnitude of the two types of DICs varies across models and may not favor the same one. In the current study, we focus on the marginal DIC, as it has been implemented in M*plus* and has been recommended in hierarchical Bayesian models for its ability to evaluate a model's generalizability beyond the observed individuals (Merkle et al., 2019).

where $\boldsymbol{y}$ is the observed data, and $\boldsymbol{\theta}$ is a vector of model parameters. We note that $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior, which is multiplied by the probability of the replicated data given the model parameters ($p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})$). This equation can be expanded to

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{12}$$

where the posterior is replaced by the product of the likelihood ($p(\boldsymbol{y}|\boldsymbol{\theta})$) and the prior ($p(\boldsymbol{\theta})$).

MCMC is used to draw from the posterior predictive distribution, and a replicated dataset is obtained from each draw of $\boldsymbol{\theta}$ using $p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})$. Next, a discrepancy function is used to make a comparison between the observed and replicated data. This discrepancy function tests a model M_0 against an unrestricted mean and covariance matrix model, M_1 (Muthén, 2010). There are many forms that the discrepancy function can take on (Gelman et al., 1996), but the classic likelihood ratio test (LRT) is commonly implemented as follows, such as in M*plus* (Asparouhov & Muthén, 2010a):

$$\begin{aligned} F_{ML} &= D \\ &= \frac{n}{2}(\log|\Sigma| + Tr(\Sigma^{-1}(\boldsymbol{CV} + (\mu - \bar{x})(\mu - \bar{x}))) \\ &\quad - \log|\boldsymbol{CV}| - q), \end{aligned} \tag{13}$$

where n is the sample size, Σ is the model implied covariance matrix, $\boldsymbol{CV}$ is the sample covariance matrix, μ is the model implied mean, $\bar{x}$ is the sample mean, and q is the number of observed variables in the model (Asparouhov & Muthén, 2021; Scheines et al., 1999). Estimates for μ_s and Σ_s are computed based on the M_0 model parameter estimates at each s iteration in the chain. Next, the discrepancy function based on observed data is computed as $D_s^{obs} = D(\bar{x}, \boldsymbol{CV}, \mu_s, \Sigma_s)$.

The next step deals with the replicated data. A replicated dataset is generated at each iteration of the Markov chain for the M_0 model. A discrepancy function is formed for the replicated data as $D_s^{rep} = D(\bar{x}_s, \boldsymbol{CV}_s, \mu_s, \Sigma_s)$, which is based on the sample mean ($\bar{x}_s$) and covariance matrix ($\boldsymbol{CV}_s$) for the replicated data. Next, a reference distribution $P_{reference}$ is derived from the joint distribution of $\boldsymbol{y}^{rep}$ and $\boldsymbol{\theta}$ as follows:

$$P_{reference}(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) = p(\boldsymbol{y}^{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}). \tag{14}$$

The reference distribution $P_{reference}$ is then used to evaluate the discrepancy $D(\boldsymbol{y}, \boldsymbol{\theta})$ using a tail probability akin to the classic p-value (Congdon, 2007; Scheines, Hoijtink, & Boomsma, 1999). This tail probability is called a posterior predictive p-value (PPP-value):

$$\text{PPP-value}(\boldsymbol{y}) = P_{reference}\left[D(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) > D(\boldsymbol{y}, \boldsymbol{\theta})|\boldsymbol{y}\right], \tag{15}$$

which can also be written as follows:

$$\text{PPP-value} = p(D^{rep} > D^{obs}) \approx \frac{1}{S}\sum_{s=1}^{S} \delta_s, \tag{16}$$

where S is the number of iterations in the Markov chain, and $\delta_s = 1$ if $D_s^{rep} > D_s^{obs}$ and 0 otherwise (Asparouhov & Muthén, 2021). This process involves computing $D(\boldsymbol{y}_s^{rep}, \boldsymbol{\theta}_s)$ and $D(\boldsymbol{y}, \boldsymbol{\theta}_s)$, and next the proportion of s samples where $D(\boldsymbol{y}_s^{rep}, \boldsymbol{\theta}_s)$ exceeds $D(\boldsymbol{y}, \boldsymbol{\theta}_s)$ is computed (Congdon, 2007). Extreme low PPP-values can indicate that there may be model misspecification (Asparouhov & Muthén, 2010b; Cain & Zhang, 2019). We note that it is not recommended to use a standard frequentist p-value cutoff (e.g., 0.05) when interpreting the PPP-value because PPP-value tends to be conservative (see, e.g., Robins et al., 2000) and not reflected well by the standard frequentist cutoff. However, this is still a common cutoff that is implemented, so we consider implications of using it in our investigation.

3.3. Bayesian Approximate Fit Indices

It has become standard reporting to include approximate fit indices within frequentist structural equation modeling. One benefit to these approximate fit measures is that they can help to identify models that fit the data in an approximate (but substantively accurate) sense. We highlight three approximate fit indices that have recently been adopted into the Bayesian framework: the root mean square error of approximation (RMSEA; Steiger & Lind, 1980; Steiger, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), and the comparative fit index (CFI; Bentler, 1990).

3.3.1. Bayesian Root Mean Square Error of Approximation (BRMSEA)

The RMSEA is an absolute index used to assess "badness-of-fit," and it can be converted for use within Bayesian statistics. Specifically, the BRMSEA is computed at each iteration of the Markov chain as follows:

$$BRMSEA_s = \sqrt{\max\left[0, \frac{D_s^{obs} - p^*}{(p^* - pD)n}\right]}, \tag{17}$$

where D^{obs} is the observed data discrepancy function, s is a given iteration in the Markov chain, n is the sample size, p^* is the number of parameters in the target model, and pD is typically close to the number of parameters in the M_0 model when no informative priors are specified (Garnier-Villarreal & Jorgensen, 2020). The BRMSEA captures model misfit through the rescaled discrepancy at iteration s. A χ^2-based distribution of realized values can be constructed based on BRMSEA (Garnier-Villarreal & Jorgensen, 2020), which ties this index to the PPMC procedure described above.

3.3.2. Bayesian Tucker-Lewis Index (BTLI)

The BTLI (Asparouhov & Muthén, 2021; Depaoli, 2021; Garnier-Villarreal & Jorgensen, 2020) is based on the deviance, which we assume is evaluated at the posterior mean here (but it can take on alternative forms). The BTLI can be written as follows:

$$BTLI_s = \frac{(D_{b,s}^{obs} - pD_b)/(p^* - pD_b) - (D_{t,s}^{obs} - pD_t)/(p^* - pD_t)}{(D_{b,s}^{obs} - pD_b)/(p^* - pD_b) - 1}, \tag{18}$$

where D^{obs} is the observed data discrepancy function, b is the baseline model with no covariance structure, t is the target model, p^* is the number of parameters in the model, and s is a given iteration in the Markov chain.

3.3.3. Bayesian Comparative Fit Index (BCFI)

The BCFI (Asparouhov & Muthén, 2021; Depaoli, 2021; Garnier-Villarreal & Jorgensen, 2020) can take on different versions, but we present a version where the Markov chain is rescaled using pD, indicating that the expectation is equal to the deviance evaluated at the posterior mean. The index can be written as follows:

$$BCFI_s = 1 - \frac{D^{obs}_{t,s} - p^*}{D^{obs}_{b,s} - p^*}, \tag{19}$$

where D^{obs} is the observed data discrepancy function, b denotes the baseline model with no covariance structure, t denotes the target model, s is a given iteration in the Markov chain, and p^* is the number of parameters in the target model.

3.4. Performance of Model Fit and Selection Indices

We just covered several indices that can be used to detect model misspecification for Bayesian modeling. The performance of these indices in the literature are highlighted as follows.

In LGMs, a typical example is fitting a linear model even though the true trajectory in the population is quadratic or piecewise. Research has been done in the frequentist framework to examine how sensitive the fit indices are to this type of misspecification. Yu (2002) simulated data from two quadratic models with either five or eight measurement occasions, and varied sample sizes from 100 to 1000. The analysis model was misspecified by dropping the quadratic form. It was found that when using a cutoff value of 0.95 (Hu & Bentler, 1999), CFI and TLI performed well for both the correct and the misspecified models with five measurement occasions and $n \geq 250$. With eight measurement occasions, the CFI and TLI could capture 100% of misspecified models even with $n = 100$. Using a cutoff value of 0.06 (Hu & Bentler, 1999), RMSEA worked better in capturing misspecified models but were more likely to reject the correct model for the five-occasion model with $n = 250$. Leite and Stapleton (2011) generated data from different nonlinear growth trajectories, including linear-linear piecewise, with six measurement occasions, and fit a linear model to them. Different sample sizes and degrees of misspesification were examined. They found that CFI and TLI tended to retain most of the misspecified models when the cutoff value of 0.95 was used, regardless of sample size or severity of misspesification. RMSEA, on the other hand, was able to detect misspecification 90% of the time when the cutoff value of 0.06 was used.

For piecewise LGMs, another typical form of misspecification is knot misplacement. Ning and Luo (2017) conducted a simulation study to investigate the impact of misplacement of the knot on model fit. The data were generated based on a linear-linear piecewise model with seven measurement occasions. The knot was placed at one of the following four occasions in the data generation process: 3, 3.5, 4, and 5. In the analysis model, the knot was fixed at the third occasion, which created four scenarios: no misspecification, misspecification of 0.5, 1, or 2 occasions. They found that when data followed the normal distribution, RMSEA, CFI and TLI all performed well with the true model. However, none of them were sensitive to the knot misplacement, regardless of sample size or severity of misspecification. When data were skewed, RMSEA became more sensitive as the severity of misspecification increased, while CFI and TLI still failed to capture this type of misspecification.

Only a few studies have focused on the performance of Bayesian model fit and model selection indices. Comparisons between PPP-value and approximate fit indices in confirmatory factor analysis models can be found in Winter (2021). She concluded that the PPP-value was more sensitive to misspecification (e.g., reducing number of factors; fixing cross-loading to 0) than BCFI and BTLI. Asparouhov et al. (2015) recommended using DIC over BIC when comparing Bayesian structural equation models. Research on the performance of these indices in Bayesian LGMs has been even more limited. Only until recently, one study (Winter & Depaoli, 2022) systematically investigated how Bayesian model fit indices, PPP-value, BRMSEA, BCFI, BTLI, as well as the model selection indices, BIC and DIC, were sensitive to three types of misspecification (i.e., constraining measurement errors to be equal; fixing variance of quadratic slope at 0; and completely dropping quadratic slope) in a latent quadratic model. They adopted the commonly used cutoff values for good fit: PPP-value $>.05$, BRMSEA $<.06$, and BCFI and BTLI $>.95$. They also assessed the impact of sample size, missing data and prior distribution on the performance of these indices. The PPP-value showed the best performance, except for $n = 50$. BRMSEA worked adequately well, unless when $n = 50$ or 100, or diverging (i.e., wrong) priors were used. BCFI and BTLI were the least sensitive to model misspecification, even with a large sample size such as 500. Neither BIC nor DIC preferred the models with misspecification in quadratic slope, however, both of them tended to favor the model with constrained measurement errors over the correct model. As sample size increased, the DIC became more likely to select the correct model, while the selection of BIC was not impacted by sample size, at least examined in this study. Prior specifications did not influence the performance of the BIC or DIC. The negative impact of missing data on all the indices was only observed when there were 50% of missing values on four measurement occasions. This body of research points toward the need for reliable measures that can detect a variety of model misspecifications. However, there is still a complete lack of information about performance of these indices for Bayesian piecewise LGMs. In what follows, we detail a simulation study aimed at uncovering performance and ability to detect specification errors in this modeling context.

4. Simulation Design

We conducted a Monte Carlo simulation study to evaluate the performance of several Bayesian (approximate) model fit indices in piecewise LGMs in detecting misspecification in knot placement and examine the impact of priors on their performance. In addition, we examined the performance of two Bayesian model comparison indices (the BIC and DIC).

4.1. Population Models

We focused on a linear-linear Bayesian piecewise LGM with seven measurement occasions and a slope change at the fourth occasion (Figure 1). The population values were adopted from those in Ning and Luo (2017).

4.2. Magnitude of the Change in Slope

Based on Kwok et al. (2010), we considered 3 levels in the magnitude of change in the linear slope. We first fixed the slope of the first segment of growth at 0.5, then the second slope was set at 0.56, 0.66, or 0.75 to reflect "small," "medium," and "large" changes. These trajectories are pictured in Figure 2. The changes between the two linear slopes were computed as the product of the standardized effect size (0.2, 0.5 and 0.8) based on Cohen (1988) and the standard deviation of the first linear slope ($\sqrt{0.1}$), following the Raudenbush and Xiao-Feng (2001) effect size equation.

4.3. Knot Placement

We manipulated knot placement at 4 levels to reflect common scenarios/mistakes seen in applied studies. In the analysis model, the knot was placed at the true location (correct model), one time point before the true location, one time point after the true location, or completely ignored (i.e., no knot; the two-piece model then reduced to a single-piece linear model). Overall, there were one correctly specified model and three misspecified models (regarding knot location). The two factors, knot placement and magnitude of the change in the growth rate, defined the levels of misspecification.

4.4. Sample Size

Sample size is a consistent factor tied to the performance of model fit measures implemented within SEM, including Bayesian SEM (see e.g., Garnier-Villarreal & Jorgensen, 2020; Shi et al., 2019). As a result, it is important to examine the performance of these indices for piecewise LGMs across a range of sample sizes in order to determine an accurate picture of their performance. We included 5 levels of sample size: 30, 75, 150, 300 and 500, which represent a reasonable range of sample sizes that are commonly seen in practice. The smallest sample size $n = 30$ was chosen following previous simulation studies (e.g., Kwok et al., 2010; Ferron et al., 2002; Keselman et al., 1998) in order to explore the impact of different prior distributions in extreme circumstances.

4.5. Prior Specification

We examined the impact of different prior specifications in order to gain a complete understanding of how priors impact the ability of the model fit and selection indices to detect model misspecification. We examined the following prior conditions for the means of the latent intercept and slopes: (1) diffuse; (2) informative accurate; (3) informative inaccurate; (4) weakly informative accurate; and (5) weakly informative inaccurate. All these priors followed the normal distribution, with a mean hyperparameter μ and a variance hyperparameter σ^2. For the diffuse conditions, we used the default prior specification in M*plus*, in which $\mu = 0$, and $\sigma^2 = 10^{10}$. The informative priors were those with high precision, in which the variance hyperparameters were set at small values. In contrast, the weakly informative priors were those with large variance hyperparameters (but still much smaller than the diffuse condition). We also manipulated the accuracy of the priors by centering them at the true population values (accurate) or shifting them upward (inaccurate). Specifically, the accurate informative priors were centered at the true population means, and the variance hyperparameters were set at 0.1 times the true population means. The inaccurate informative priors had the same variance hyperparameters as the accurate priors, while their means were shifted upward by 3 times the square root of the variance hyperparameters (3σ). We specified the weakly informative priors (accurate and inaccurate) in a similar manner. The only difference was that for both weakly informative priors, their variance hyperparameters were set at 0.5 times the true population means. Table 1 shows the values of the hyperparameters for all these priors. The M*plus* default priors were used for all other parameters.

4.6. Data Generation and Bayesian Estimation

We used M*plus* version 8.6 (Muthén & Muthén, 1998–2017) for data generation and estimation via the Bayesian framework. For Bayesian analyses, we implemented the Gibbs sampler with 2 chains each consisting of 10,000 iterations.

Table 1. Normal distribution settings for latent growth factor means: hyperparameter values for simulation conditions.

Factor	Mean Hyperparameter	Variance Hyperparameter
Informative prior settings		
Accurate location		
Intercept	2.500	0.250
Slope 1	0.500	0.050
Slope 2	0.560	0.056
Inaccurate location		
Intercept	4.000	0.250
Slope 1	1.171	0.050
Slope 2	1.270	0.056
Weakly informative prior settings		
Accurate location		
Intercept	2.500	1.250
Slope 1	0.500	0.250
Slope 2	0.560	0.280
Inaccurate location		
Intercept	4.000	1.250
Slope 1	1.171	0.250
Slope 2	1.270	0.280

We found that this number of iterations was enough after testing several chain lengths and visually checking convergence (trace plots were uploaded in the online material). The first half of the iterations was discarded as the burn-in phase of the chain. Convergence was monitored using the $\hat{R}$ convergence diagnostic (Vehtari et al., 2019).

4.7. Outcomes of Interest

We first examined the values of the model fit indices, PPP, BRMSEA, BCFI, and BTLI, in all conditions to assess how they performed in detecting different levels of misspecification, and how their performance was impacted by prior specification and sample size. In addition, we used 90% credible intervals (CIs) of the approximate fit indices (Asparouhov & Muthén, 2021) as an assessment to evaluate their sensitivity to misspecification. If the entire 90% CI fell below .06 for BRMSEA, or above .95 for BCFI or BTLI, then the model fit is "good." If the entire 90% CI fell on the other side of the cutoff values, it indicates a "poor" model fit. If the cutoff value is within the 90% CI, the model fit is "inconclusive." We then computed the proportion of replications in which the model fit was "good," "inconclusive," or "poor."[3] For the correctly specified model, we expect to see a high proportion of good fit. Similarly, for the misspecified models, the higher proportion of poor fit an index produced, the more sensitive it was to misspecification. For the model selection indices, BIC and DIC, we assessed the model selection rates across various conditions to examine how often they favored the correct model over a misspecified model. The difference between the indices from the two models being compared was also computed.

5. Simulation Results

The BCFI and BTLI results were very similar, so we will only be reporting the BCFI results for the sake of space.[4] To assess chain convergence, we extracted $\hat{R}$, and we used cutoff criteria of $\hat{R}<1.05$ as an indicator of convergence. On average, 94.51% of all replications had a maximum $\hat{R} < 1.05$. We included all converged replications in the results that follow.

5.1. Model Fit: Using Index Values as an Assessment

Model fit measures are commonly used as tools for identifying model misspecification. As misspecification worsens, it would be expected that the model fit indices would reflect this worsening. Figures 3–5 present boxplots of the index values across all conditions of correct and misspecified models. Within each of these figures, the columns represent the slope rate of the second segment for the piecewise model, and the rows represent sample size conditions (smallest on top and largest sample size on bottom). The y-axis represents the model fit index value. Finally, the x-axis represents the different prior conditions examined, and the boxes represent model misspecification (correct and incorrect).

5.1.1. PPP

Results for the PPP-value are in Figure 3, and each plot includes two horizontal lines. The solid horizontal line represents the PPP-value of 0.5, which would indicate optimal model fit (Asparouhov & Muthén, 2010). The dashed horizontal line is set at PPP = 0.05 to showcase a common cut-point that applied researchers may implement to determine model misfit (Asparouhov & Muthén, 2010; Zyphur & Oswald, 2015). We will use this cut-point as an illustration for interpreting the results, but it is important to note that there is no single cutoff value that should be used across all Bayesian modeling situations implementing the PPP procedure–we use this value for convenience in interpreting results. If PPP is working as expected, then correctly specified models should hover closer to the PPP = 0.5 value (solid horizontal line) and misspecified models should hover around or below the PPP = 0.05 value (dashed horizontal line).

In order to become oriented with the PPP findings, start by looking at the top-left corner for the slope of 0.56 and $n = 30$. The correct ("true location") model and the two models with the knot placed one time-point before and after the true location (i.e., the three lightest shaded boxes) all look comparable to one another. The correct model (lightest shade) is not hovering around the expected value of PPP = 0.5. In fact, the ignored knot condition (darkest shade) has more mass surrounding the 0.5 line than the other three. In addition, none of the cells (across any of the prior conditions) were flagged as representing misfit with the dashed cutoff value. These results indicate that the PPP cannot properly classify the smallest slope and smallest sample size condition with respect to the cutoffs typically implemented. Moving over to the other extreme cell, located in the bottom-right corner, results appear as expected. The correctly specified model with the true knot location is hovering over the PPP = 0.5 line perfectly across prior conditions. In addition, the ignored knot condition is clearly below the PPP = 0.05 line, and the inaccurate knot location boxes (misspecified models, but not as egregious as ignoring a knot altogether) hover over that line to a large degree.

Zooming out, there are many general findings that can be pinpointed here. As sample size increases from small to large (i.e., looking down the rows), the PPP is better able to distinguish between correct and misspecified models, pinpointing the ignored knot condition as containing the largest degree of misfit. In looking at the columns, the size of the slope also plays a role in overall findings. The PPP was better able to identify misfit in the knots when the second

[3] Although we implemented a 90% CI here, it is important to note that other interval widths could have also been implemented. It is possible that substantive conclusions would differ with the implementation of a different interval width (e.g., more inconclusive decisions may be made if wider widths are used). However, we selected 90% to be consistent with the defaults in M*plus*. We felt that this would be the most informative setting because it maintains consistency with how the methods will likely be applied in the literature.

[4] For full results, please see the OSF page for this project: https://osf.io/myrds/

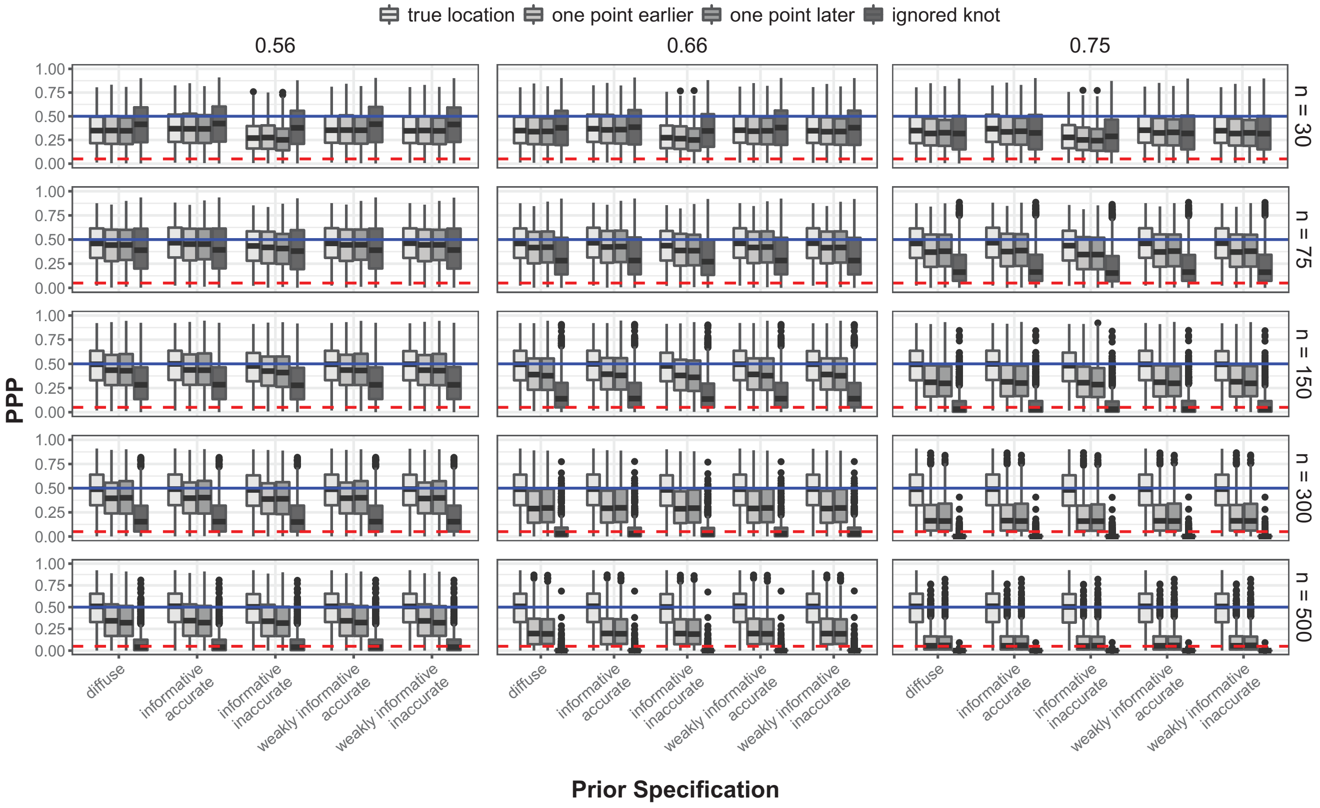

Figure 3. PPP across simulation conditions.

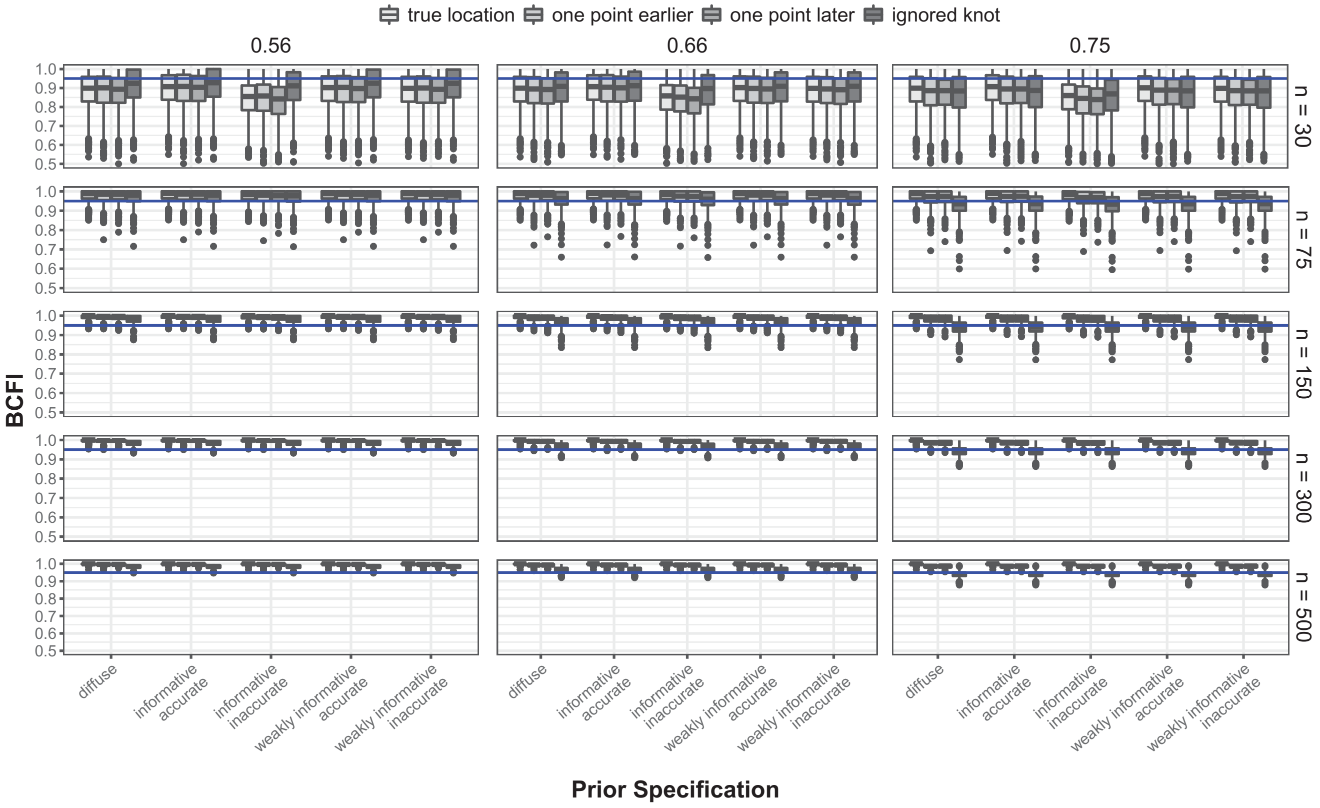

Figure 4. BCFI across simulation conditions. Note that the y-axis has been truncated at 0.5 to aid in interpreting the patterns across all cells in the figure.

segment's slope was larger (right column) as compared to a smaller slope (left column). Regarding prior specifications, results across the priors are largely comparable within each subplot, with one interesting exception. Focusing on the "informative inaccurate" prior setting (middle prior condition in each subplot), it is clear that this prior is pulling the

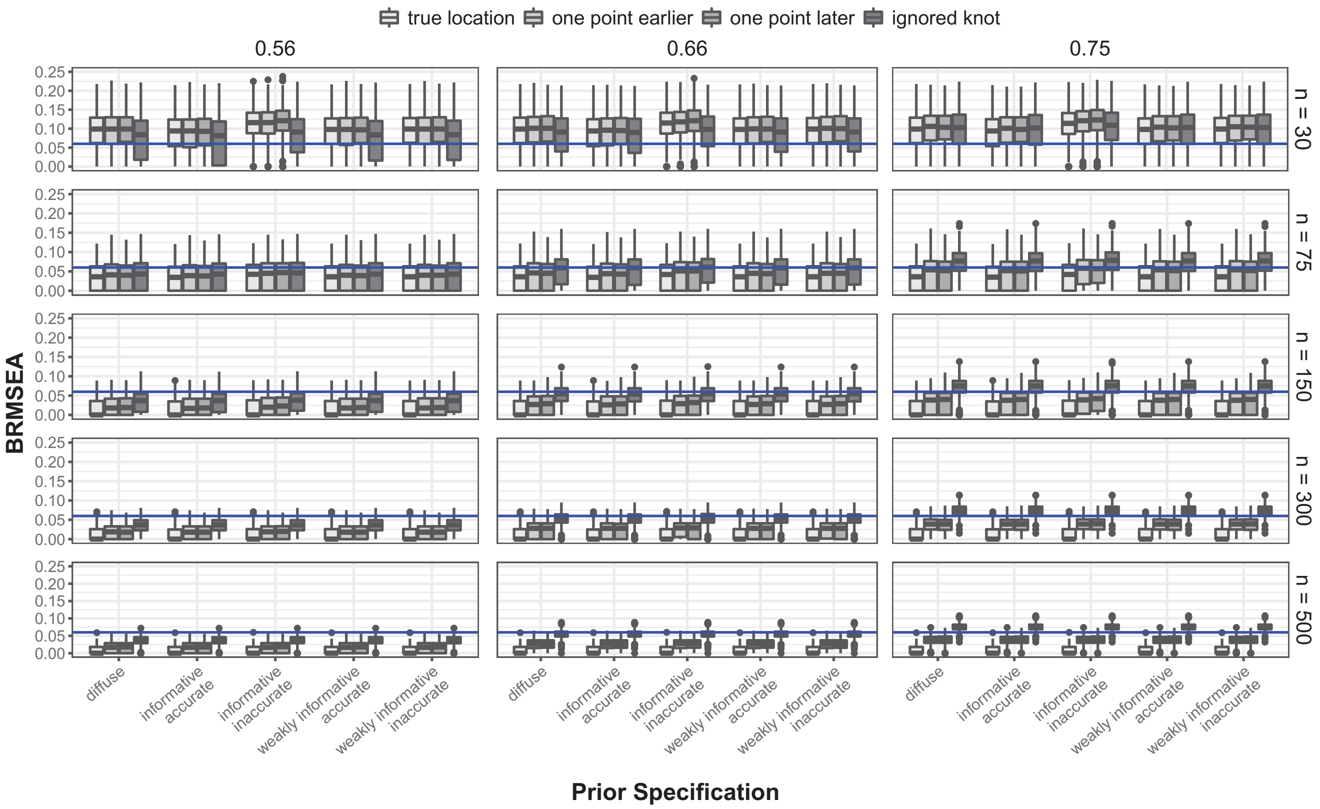

Figure 5. BRMSEA across simulation conditions.

PPP downward, especially under smaller sample sizes. In other words, the inaccurate (but precise) priors add to the degree of misfit for all small sample size conditions. Otherwise, there is no appreciable influence of prior specification on the PPP results. For sample sizes under 500, it is more difficult for the PPP to identify misfit according to a strict cutoff (e.g., PPP < 0.05 represents misfit).

5.1.2. BCFI (and BTLI)

Figure 4 contains results for the BCFI. The figure is read much the same way as with the PPP, with one notable difference. We have added a horizontal line at BCFI $= 0.95$ to illustrate a common cutoff that applied researchers use to determine when a model should be rejected or not (values < 0.95 would indicate misfit under this common cutoff; Garnier-Villarreal & Jorgensen, 2020; Asparouhov & Muthén, 2021). An additional point to note is that the y-axis has been zoomed in so that the lowest number plotted on this axis is 0.5 (i.e., we did not start the y-axis at a value of 0). Notably, some of the most extreme outliers for $n = 30$ were cut off with this truncated y-axis. However, this scale allows for a more detailed (zoomed in) viewed of the results for the larger sample sizes. Overall, there is a clear sample size effect in this figure. Using a strict cutoff value of 0.95 illustrates that none of the models (even the correctly specified model) reliably met that cutoff for $n = 30$ and, for sample sizes $n = 75$ and above, the cutoff is not able to reliably identify mis-specifications; with the exception of the bottom right box (representing the most "extreme" levels of misspecification with a slope of 0.75 and the largest sample size). Overall, the BCFI (and BTLI–pictured in the online material) were not informative about misfit with the use of a 0.95 cutoff value. In addition, there did not appear to be any meaningful differences in the prior specifications implemented for the BCFI.

One interesting element to note, is that it appears that under some conditions BCFI (and BTLI) can be used for *model selection* purposes. Take, for example, the sample sizes > 30 and look at the four adjacent boxplots for any prior condition. The median BCFI is always larger, even if only slightly so, for the true model. In turn, the lowest median BCFI is always associated with the condition ignoring the knot altogether. This implies that the BCFI can be used within a specific prior and sample size condition to help determine the true model. The results are most exaggerated for the largest sample size and the largest slope condition (i.e., bottom right plot in the figure). We expand on this concept of using the indices for model selection in the Discussion section.

5.1.3. BRMSEA

Figure 5 shows the results for the BRMSEA. In this figure, we added a horizontal line at 0.06 to showcase when a model would be rejected if this cutoff point was used (values > 0.95 indicate misfit; Garnier-Villarreal & Jorgensen, 2020; Asparouhov & Muthén, 2021). Overall, the BRMSEA was not able to identify correct versus incorrect models at $n = 30$. As sample sizes increased, the ability to detect the correct model (lightest box) from the model with the ignored knot (darkest box) became clearer. However, using the cutoff value, the misspecified knot location conditions (middle two boxes within each group of boxes) were not identified as misspecifications. Regarding results across columns, the slope condition of 0.75 (right column) was the only one to reliably identify misfit for the "ignored knot" condition. The results indicate that the smaller changes in slope (left and middle columns) cannot be identified by the BRMSEA as misfit (even with explicit knot misspecifications). Just as with the previous indices, there were no notable differences in the priors specified.

5.2. Model Fit: Using 90% Credible Intervals as an Assessment

The BCFI, BTLI, and BRMSEA all produce credible intervals (CIs) that can be used as another mechanism for assessing model fit results. One way that these approximate fit CIs can be used is to classify fit as being "good," "inconclusive," or "poor." Examining the CI can potentially be more informative (or flexible) as compared to a single cutoff value. Figures 6 and 7 contain the CI results for the BCFI and BRMSEA, respectively; BTLI results can be found online and were comparable to BCFI. These two figures contain stacked bars, which represent the proportion of replications that resulted in "good," "inconclusive," or "poor" model fit. The figures are set up as follows. Columns represent the slope of the post-knot segment and rows represent sample size. The x-axis represents the specification of the knot location: true location, one point earlier, one point later, or when the knot is completely ignored. The y-axis represents the proportion of replications falling in either "good," "inconclusive," or "poor" classification based on the 90% CI. Specifically, the "good" category is defined by a 90% CI that is entirely in the range of BCFI ≥ 0.95 or BRMSEA ≤ 0.06. The "poor" category is defined by an interval that is entirely in the range of BCFI < 0.95 or BRMSEA > 0.06 (see, e.g., Asparouhov & Muthén, 2021; Winter & Depaoli, 2022). The "inconclusive" category has an interval that contains the cutoff value, where part of the interval is above and the rest of the interval is below that cutoff value. Given that there was very little difference in results across different prior settings, the results were collapsed across prior specifications.

The results were quite similar across BCFI and BRMSEA, so we will focus on narrating BCFI in Figure 6. Looking across sample sizes (rows), there are different patterns of performance. For $n = 30$, the great majority of replications had results that were either "inconclusive" or "poor," with little-to-no replications classified as "good." Sample sizes needed to be ≥ 75 in order for CIs to fall within the "good" category. As sample sizes increased beyond $n = 75$, the proportion of replications falling in the "good" category also increased. At the largest sample size of $n = 500$, and the smallest slope of 0.56 (bottom left corner plot), almost all of the replications produced results indicating "good" fit according to the 90% CI. However, as the slope value increased to 0.75 (bottom right corner plot), the ignored knot condition indicated a high degree of misfit with most replications indicating "poor" fit according to the 90% CI. Overall, findings according to the CI indicate that very few replications would be classified as "good" fit at sample sizes $n \leq 75$, even for the true model. As sample size increased, however, even the misspecified

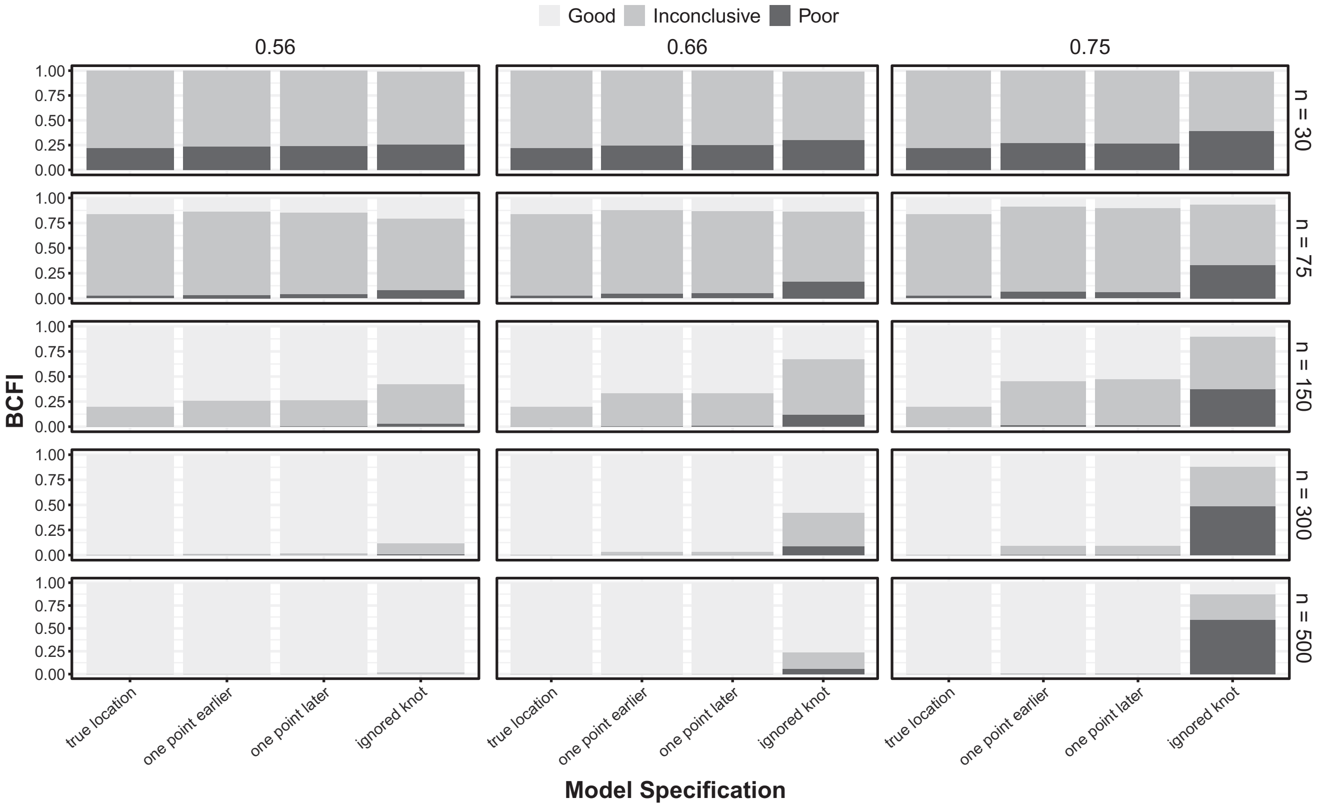

Figure 6. BCFI 90% credible interval rejection rates.

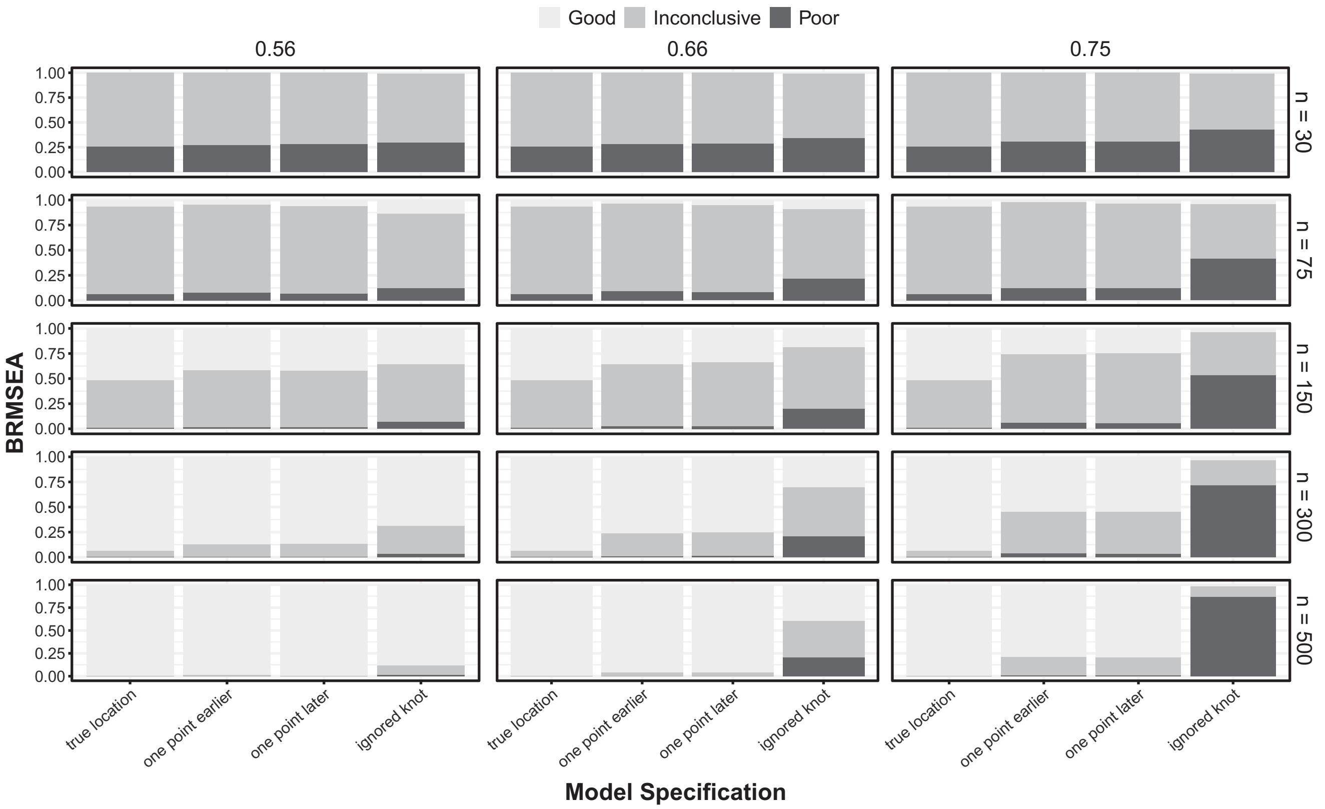

Figure 7. BRMSEA 90% credible interval rejection rates.

models with a knot placed at the wrong location were classified as well-fitting models for most replications. Ignoring the knot completely was classified to a larger extent as misfit (or "inconclusive"), especially under larger sample sizes and the two larger slope conditions.

5.3. Model Comparison Using the BIC and DIC

Tables 2 and 3 contain model comparison results for the BIC and DIC, respectively. The main horizontal blocks of these tables contain information for each of the three post-knot slope conditions: 0.56, 0.66, and 0.75. Results are presented by sample size (rows) and prior specification (columns). The values in the table showcase two types of information: selection rates and degree of difference in the information criteria. The numbers in the tables represent the percent of replications where the true model was favored by the index as compared to one of the three following misspecified models: knot location was one time-point earlier than the true model, knot location was one time-point later than the true model, or the knot was completely ignored. Higher percentages in the table represent conditions where the index was able to properly identify the true model at a higher rate. The second type of information in the table showcases the degree of the difference in information criteria. It is not informative enough to simply say that Model 1 had a smaller BIC value as compared to Model 2. The degree of separation between the index values is also an important piece of information to consider. For example, a researcher would want to know if the BIC comparisons were Model 1 BIC = 10 and Model 2 BIC = 10.1, versus if the BIC comparisons were Model 1 BIC = 10 and Model 2 BIC = 1000. These two scenarios represent differing magnitudes of difference between the BIC values for competing models. In order to highlight this layer of results, we have bolded selection rate values where the average difference between the information criterion values was greater than 5. Non-bold values indicate that selection rates were based off of an average point differential that was less than 5 (i.e., the selection was based on a narrower gap between the information criterion values).

Results for the BIC, found in Table 2, presented a few important findings that we will highlight here. The most striking results surround the patterns of bold values, which showcase a larger discrepancy between the BIC values yielded for the competing models. In comparing results across the columns, the true model and ignored-knot model (right column) had larger discrepancies in the BIC values as compared to the other two columns (true model versus a model with an incorrect knot placement). The cells with the largest discrepancies and the largest selection rates of the true model are across the bottom of the table. Specifically, the largest sample sizes ($n \geq 300$) for the largest slope value (0.75) had consistently larger selection rates that are in bold font. The moderate slope of 0.66 showed similar patterns for $n \geq 300$ as well. These cells represent the conditions where the BIC was most decisive in properly detecting the true model when compared to a misspecified model. Overall, the BIC did not pinpoint specification errors accurately for smaller sample sizes (<300), even with severe misspecifications (e.g., bottom right panel, representing slope = 0.75 and ignored knot).

Table 2. BIC selection rates: comparing the true model to three misspecified models.

	True model vs. one point earlier					True model vs. one point later					True model vs ignored knot				
n	DIF	I-A	I-INA	WI-A	WI-INA	DIF	I-A	I-INA	WI-A	WI-INA	DIF	I-A	I-INA	WI-A	WI-INA
Slope = 0.56															
30	53.8	54.2	49.3	53.7	53.6	53.1	53.0	67.1	53.0	53.9	**0.4**	**0.4**	**0.1**	**0.4**	**0.4**
75	61.3	61.4	60.1	61.4	61.1	57.3	56.9	65.0	56.9	57.8	**1.4**	**1.4**	**1.3**	**1.4**	**1.4**
150	67.8	67.8	66.9	67.8	67.8	65.3	65.3	69.1	65.3	65.6	**4.7**	**4.7**	**4.6**	**4.7**	**4.7**
300	76.6	76.6	76.6	76.6	76.6	75.3	75.3	76.4	75.3	75.5	**12.6**	**12.6**	**12.2**	**12.6**	**12.6**
500	82.5	82.5	82.6	82.5	82.5	84.2	84.2	84.7	84.2	84.2	33.3	33.3	33.1	33.3	33.3
Slope = 0.66															
30	55.5	56.5	56.0	55.4	55.6	57.1	57.0	68.4	57.2	58.4	**1.5**	**1.5**	**0.8**	**1.5**	**1.5**
75	68.7	68.8	68.1	68.7	68.6	65.9	65.8	70.5	65.9	65.9	**5.6**	**5.6**	**4.8**	**5.6**	**5.6**
150	76.0	76.1	76.1	76.0	76.0	75.1	75.0	77.9	75.1	75.2	**20.7**	**20.7**	**20.3**	**20.7**	**20.7**
300	**85.5**	**85.5**	**85.5**	**85.5**	**85.5**	**84.2**	**84.2**	**84.8**	**84.2**	**84.3**	55.9	55.9	55.7	55.9	55.9
500	**92.5**	**92.5**	**92.6**	**92.5**	**92.6**	**92.8**	**92.8**	**92.9**	**92.8**	**92.8**	**89.3**	**89.3**	**89.1**	**89.3**	**89.3**
Slope = 0.75															
30	60.8	61.9	63.1	61.0	61.3	61.5	61.7	71.4	61.4	62.9	**4.8**	**4.9**	**3.0**	**4.8**	**4.8**
75	76.5	76.6	77.1	76.5	76.4	74.4	74.3	77.6	74.4	74.5	**22.6**	**22.8**	**20.7**	**22.6**	**22.5**
150	83.5	83.5	84.0	83.5	83.5	**86.8**	**86.8**	**87.2**	**86.8**	**86.8**	57.8	57.9	57.1	57.8	57.7
300	**93.5**	**93.5**	**93.5**	**93.5**	**93.5**	**94.2**	**94.2**	**94.2**	**94.2**	**94.2**	**96.1**	**96.1**	**95.8**	**96.1**	**96.1**
500	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**98.2**	**98.2**	**98.2**	**98.2**	**98.2**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**

Note. n is the sample size. DIF is the diffuse prior. I-A is the informative-accurate prior. I-INA is the informative-inaccurate prior. WI-A is the weakly informative-accurate prior. WI-INA is the weakly informative-inaccurate prior. Numbers in the table represent selection rates in terms of the percentage of replications where the true model was favored over the misspecified model. The degree of difference in information criterion values is captured through bold values. For interpretation purposes, we bolded selection rates for conditions where the difference in information criteria values between the true model and misspecified model was greater than 5 points.

Table 3. DIC selection rates: comparing the true model to three misspecified models.

	True model vs. one point earlier					True model vs. one point later					True model vs ignored knot				
n	DIF	I-A	I-INA	WI-A	WI-INA	DIF	I-A	I-INA	WI-A	WI-INA	DIF	I-A	I-INA	WI-A	WI-INA
Slope = 0.56															
30	52.7	50.1	46.2	51.0	51.2	53.1	50.7	63.7	52.1	53.7	17.5	22.6	11.0	19.0	17.8
75	61.7	59.5	57.7	60.8	60.5	53.0	52.0	62.2	52.5	53.4	31.2	33.0	29.9	31.7	31.7
150	65.8	65.2	64.4	65.6	65.3	65.2	64.4	69.5	64.7	65.3	59.8	61.2	58.3	60.0	59.8
300	76.2	75.8	75.9	76.1	76.1	74.9	74.8	76.3	74.8	74.9	**84.2**	**84.4**	**83.9**	**84.2**	**84.2**
500	81.1	81.1	81.4	81.1	81.2	84.4	84.4	84.7	84.4	84.4	**96.8**	**96.8**	**96.5**	**96.8**	**96.8**
Slope = 0.66															
30	55.8	52.4	52.7	54.1	54.4	56.7	55.5	64.9	56.4	57.3	28.5	34.6	20.5	30.0	29.0
75	68.5	67.3	67.2	68.2	68.0	63.6	63.4	68.3	63.6	64.1	53.4	55.8	52.3	54.0	54.0
150	75.7	75.2	75.2	75.4	75.4	76.3	75.9	77.9	76.3	76.5	**85.5**	**86.4**	**84.6**	**85.6**	**85.6**
300	**85.4**	**85.3**	**85.5**	**85.3**	**85.4**	**84.4**	**84.4**	**85.0**	**84.4**	**84.4**	**98.7**	**98.8**	**98.7**	**98.7**	**98.7**
500	**92.0**	**92.0**	**92.0**	**92.0**	**92.0**	**93.3**	**93.3**	**93.3**	**93.3**	**93.3**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**
Slope = 0.75															
30	61.2	60.0	61.3	60.8	60.7	62.3	60.9	68.8	61.8	62.7	44.8	49.2	36.5	45.8	45.0
75	76.4	75.5	75.8	76.3	76.3	72.7	72.7	75.9	72.8	73.0	**77.3**	**79.1**	**76.8**	**77.9**	**77.9**
150	84.0	83.5	83.8	83.7	83.7	86.4	86.4	**86.8**	86.4	86.6	**98.8**	**98.9**	**98.8**	**98.8**	**98.8**
300	**93.6**	**93.4**	**93.8**	**93.6**	**93.6**	**94.0**	**94.0**	**94.0**	**94.0**	**94.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
500	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**98.2**	**98.2**	**98.2**	**98.2**	**98.2**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

Note. *n* is the sample size. DIF is the diffuse prior. I-A is the informative-accurate prior. I-INA is the informative-inaccurate prior. WI-A is the weakly informative-accurate prior. WI-INA is the weakly informative-inaccurate prior. Numbers in the table represent selection rates in terms of the percentage of replications where the true model was favored over the misspecified model. The degree of difference in information criterion values is captured through bold values. For interpretation purposes, we bolded selection rates for conditions where the difference in information criteria values between the true model and misspecified model was greater than 5 points.

Results for the DIC can be found in Table 3. There are many aspects of the DIC results that are similar to the performance of the BIC. However, we will highlight a few instances where results differed across the two indices. Overall, the DIC selection rates appeared to be comparable, and in some cases higher than the BIC rates. For example, the right column for the true model versus the ignored-knot model, the DIC selection rates are much higher as compared to the BIC selection rates in Table 2. This difference between the BIC and DIC in the right-hand column is most notable in the smaller sample sizes ($n < 300$). In a side-by-side comparison, it appears that the DIC is more consistently able to select the true model as compared to the ignored-knot model. The performance of the DIC appears comparable to the BIC when examining the other two columns (comparing a true model to a model with misspecified knot placement). In addition, both indices show an improved ability to detect misspecification as sample sizes increase.[5]

[5] For the interested reader, we have added an additional plot in the online supplementary material. This plot showcases how the PPP and the DIC align with respect to model evaluation. Indeed, the more misspecified models (corresponding to lower PPP values) were selected less frequently by the DIC as compared to the true model (which corresponded to a higher PPP overall).

6. Discussion

We investigated the performance of various Bayesian (approximate) model fit and comparison indices via simulation. Our focus was on the piecewise LGM, which is an important model used to capture segmented growth over time. We set out to answer two main questions surrounding model and prior specification for this modeling situation. We will discuss each of these questions next and follow-up with recommendations for applied researchers, as well as future research directions.

6.1. Model Misspecification

Research Question #1: (Model Misspecification) How well do Bayesian model fit and asssessment measures detect model misspecification for piecewise LGMs?

We examined several different model fit and assessment indices implemented in the Bayesian estimation framework. As detailed in the Results section, each index carried with it different nuances regarding performance and ability to properly detect model misspecification. However, one global pattern that was uncovered in the simulation study was that the indices (speaking collectively, as a group) tended to perform better under conditions where the model specification error was more extreme (e.g., ignoring a knot when the slope has a larger shift in the second segment of the growth trajectory). It was more difficult for these indices to properly identify misspecification in knot placement as compared to the situation where the knot was ignored completely.

In examining performance for each of the indices individually, we can draw the following conclusions. Regarding model fit, it appears that the PPP is a more reliable tool as compared to the approximate fit indices explored here. The PPP was especially useful under cases of larger sample sizes, and it struggled with the smallest sample size conditions explored here. One potentially alarming finding here was that, as sample sizes increased, the approximate fit measures tended to indicate that all models (even the misspecified models) fit well. Early recommendations (Asparouhov & Muthén, 2021) indicated that the approximate fit measures were not well suited for smaller sample sizes. However, our findings corroborated that of Winter & Depaoli (2022) in that the indices (especially BCFI/BTLI) were not particularly helpful for larger sample sizes when examining the fit of an individual model. In addition, the approach using 90% CIs as an assessment tool (i.e., for BCFI, BTLI, and BRMSEA) appeared to be best under the smaller sample size conditions or conditions with the most extreme misspecification (i.e., knot ignored completely, and the largest slope condition of 0.75). The CI approach was not reliable outside of those specific conditions, and it is clear that the performance is tied to other factors (e.g., sample size) outside of model misspecification.

The current investigation also uncovered that the DIC is a more reliable index for properly identifying specification errors as compared to the BIC. However, there were important limitations uncovered that were tied to sample size and severity of specification error. The DIC performed best under the highest sample sizes (≥ 300) and under conditions where the knot was completely ignored. It was more difficult to properly detect the specification error when the knot was simply misplaced, as opposed to ignored.

6.2. Prior Specification

Research Question #2: (Prior Specification) Does prior specification (e.g., informativeness and accuracy of the prior) have an impact on the overall performance of model fit indices for piecewise LGMs?

The results obtained in this simulation study indicated that prior settings had little to do with the overall performance of the indices examined here. Specifically, diffuse priors performed comparably to subjective priors that were specified to be either accurate or inaccurate, and either informative or weakly informative. Although previous research has indicated that prior specification is an important element in properly estimating growth factor means and variances in LGMs (see e.g., Depaoli, 2013; van de Schoot et al., 2018; Depaoli et al., 2017; Smid et al., 2020), it was not a key element to detecting model misfit in this study. At least for this current investigation, prior settings do not appear to be much of a concern regarding the performance of the Bayesian model fit and assessment indices examined here.

6.3. Recommendations for Applied Researchers

The piecewise LGM can be a powerful tool, but it is possible to misinterpret or misrepresent substantive findings if the model is specified incorrectly. The current simulation study uncovered many important findings that can be used to help construct practical guidelines for using Bayesian model fit and comparison indices in this modeling context. All of the indices examined here have important limitations that were detailed in the Results section. Our overall assessment is that the PPP appears to be a more reliable fit index as compared to the approximate fit measures when a single model is being examined. Within the approximate fit measures, the BRMSEA appears to perform better than the BCFI and BTLI (which were largely comparable to one another). Regarding the model comparison indices, the DIC had more impressive selection rates as compared to the BIC, but neither of the indices worked well with smaller sample sizes. Regarding specific recommendations, we have an alternative approach that we would like to propose next.

These findings indicate that it is much more difficult to detect an inaccurate knot location as compared to a missing knot. However, the accuracy of the knot location is of utmost importance to applied researchers implementing this model. Our advice for assessing the accuracy of the knot location is to estimate competing models and implement the *full collection* of indices provided here to help determine the optimal model. One approach that might get around the limitations of the (approximate) fit measures, is to extend their use and consider them as being helpful *model comparison* tools. Specifically, when examining the results presented in Figures 3–5, a clear picture of "model comparison" emerges.

For example, when used as a model fit measure for a single model, results for the BCFI are not informative once sample sizes exceed 75. Given that this index is not recommended for smaller sample sizes (Asparouhov & Muthén, 2021), and it is not particularly helpful in a model fit sense when sample sizes are larger (see Results section), it is reasonable to critically question its overall utility in Bayesian modeling. Specifically, under larger sample sizes for which the BCFI is typically recommended, all of the models met the conventional cutoff indicating "good" model fit. The cutoff is not able to help distinguish a well-fitting model from a misspecified model. This use of the index and cutoff is unhelpful to the applied researcher looking to identify model misspecifications.

However, the index becomes informative if we stop using it as an index for assessing a single model and start examining it in a model comparison manner. Assume that a researcher fit four models to sample data as follows: a model with the knot at time-point 4 (the true model in this case), a model with the knot at time-point 3, one with the knot at time-point 5, and one without a knot. If these models were being examined as standalone models, then the BCFI is not helpful at all because all of them "fit" according to the conventional cutoff of 0.95. However, the researcher can compare the BCFIs for the four competing models, and the largest BCFI value would point toward the correct model in this instance (i.e., the model with the knot location at time-point 4). We see potential value in repurposing these approximate fit indices that are conventionally used for assessing fit of a single model into a model comparison context. It may be that the indices will better serve the goals of applied researchers if implemented in this way. However, we also recommend the use of the full collection of indices to examine consistency in the recommendations each index provides. Model fit and assessment can be viewed as a puzzle, with multiple pieces needed to help identify the optimal model. It would be unwise to rely on any single model fit or comparison measure during model selection, and that is especially true for these Bayesian indices.

6.4. Future Research Directions

Collectively, we are only at the beginning of understanding the potential role that the new Bayesian approximate fit measures can play in Bayesian modeling, as well as how they intertwine with the traditional Bayesian fit and comparison tools that have longstanding use in the field. There are many layers that still need to be investigated regarding the potential utility these indices have within latent variable modeling, and we highlight the major areas here.

The current simulation study provided a clear picture of the limitations and potential use of these indices in terms of piecewise LGMs. However, that picture may not remain consistent across all modeling contexts implementing the piecewise LGM. For example, we know that the general performance of (non-piecewise) LGMs in the Bayesian framework is tied to the complexity of the model. Linear LGMs without knots are relatively simple and straightforward to accurately estimate using Bayesian methods (see e.g., Zhang et al., 2007). However, once the model increases in complexity, for example, with mixtures or different degrees of nonlinearity included, the prior distributions have been repeatedly shown to make a large difference in overall performance of the estimation framework (see e.g., Depaoli, 2013, 2014; Lock et al., 2018).

In addition, the current investigation considered piecewise LGMs. Although these are important tools within the longitudinal latent variable modeling framework, we also want to note that there are other models that can be of potential use when examining nonlinearity within growth trajectories. Although beyond the scope of the current investigation, models such as the latent basis model (see e.g., McNeish, 2020) could also be used in a case where the researcher was not sure about the functional form of the growth trajectory. Further investigation into how well such models can compensate for misspecification of piecewise LGMs would be useful in providing a full picture of proper growth trajectory recovery.

The field simply does not yet know if that same pattern holds for these Bayesian fit indices. Specifically, the current investigation showed that priors had no measurable influence on index-performance. However, it is important to recognize that this investigation used the simplest form of piecewise LGMs. The influence of priors would likely be greater with increased model complexity (e.g., more knots, the presence of latent classes, nonlinearity in the segments, and the presence of missing data), and we do not yet know the impact of priors on the fit indices in these more complex modeling situations. This is, in our view, the biggest unknown that should be extensively examined before widespread adoption of these indices in applied piecewise LGM settings.

ORCID

Sarah Depaoli http://orcid.org/0000-0002-1277-0462
Fan Jia http://orcid.org/0000-0003-3855-532X
Ihnwhi Heo http://orcid.org/0000-0002-6123-3639

References

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis of latent variable models using Mplus. Retrieved June, 17, 2014 from https://www.statmodel.com/download/BayesAdvantages18.pdf.

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis using Mplus: Technical implementation*. Citeseer.

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling*, *28*, 1–14. https://doi.org/10.1080/10705511.2020.1764360

Asparouhov, T., & Muthén, B. (2021). *Analyzing imputed data with the Bayesian estimator in Mplus* [Tech. Rep.]. https://www.statmodel.com/download/BayesImputation.pdf

Asparouhov, T., Muthén, B. O. (2010b). *Bayesian Analysis using Mplus: Technical Implementation* [Tech. Rep.]. https://www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances:

Comments on Stromeyer et al. *Journal of Management*, *41*, 1561–1577. https://doi.org/10.1177/0149206315591075

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

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

Chou, C.-P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling*, *5*, 247–266. https://doi.org/10.1080/10705519809540104

Chung, J. M., Hutteman, R., van Aken, M. A., & Denissen, J. J. (2017). High, low, and in between: Self-esteem development from middle childhood to young adulthood. *Journal of Research in Personality*, *70*, 122–133. https://doi.org/10.1016/j.jrp.2017.07.001

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Congdon, P. (2007). *Bayesian statistical modelling*. John Wiley & Sons.

Cudeck, R., & Codd, C. L. (2012). A template for describing individual differences in longitudinal data, with application to the connection between learning and ability. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 3–24). IAP - Information Age Publishing, Inc.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*, 186–219.

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. (2021). *Bayesian structural equation modelling*. Guilford Press.

Depaoli, S., & Boyajian, J. (2014). Linear and nonlinear growth models: Describing a new Bayesian perspective. *Journal of Consulting and Clinical Psychology*, *46*, 784–802.

Depaoli, S., Rus, H. M., Clifton, J. P., van de Schoot, R., & Tiemensma, J. (2017). An introduction to Bayesian statistics in health psychology. *Health Psychology Review*, *11*, 248–264. https://doi.org/10.1080/17437199.2017.1343676

Diallo, T. M. O., Morin, A. J. S., & Parker, P. D. (2014). Statistical power of latent growth curve models to detect quadratic growth. *Behavior Research Methods*, *46*, 357–371. https://doi.org/10.3758/s13428-013-0395-1

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, *37*, 379–403.

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*, 46–70.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*, 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–760.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Harring, J. R., Strazzeri, M. M., & Blozis, S. A. (2021). Piecewise latent growth models: beyond modeling linear-linear processes. *Behavior Research Methods*, *53*, 593–608.

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement*, *73*, 537–568.

Hu, L-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. [Database] https://doi.org/10.1080/10705519909540118

Jaggars, S. S., & Xu, D. (2016). Examining the earnings trajectories of community college students using a piecewise growth curve modeling approach. *Journal of Research on Educational Effectiveness*, *9*, 445–471. https://doi.org/10.1080/19345747.2015.1116033

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics - Simulation and Computation*, *27*, 591–604. https://doi.org/10.1080/03610919808813497

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, *48*, 370–397.

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

Kroese, F. M., Adriaanse, M. A., Vinkers, C. D., Schoot, R. V. d., & De Ridder, D. T. (2013). The effectiveness of a proactive coping intervention targeting self-management in diabetes patients. *Psychology & Health*, *29*, 110–125.

Kwok, O.-M., Luo, W., & West, S. G. (2010). Using modification indexes to detect turning points in longitudinal data: A Monte Carlo study. *Structural Equation Modeling*, *17*, 216–240. https://doi.org/10.1080/10705511003659359

Lee, I. H., & Rojewski, J. W. (2009). Development of occupational aspiration prestige: A piecewise latent growth model of selected influences. *Journal of Vocational Behavior*, *75*, 82–90. https://doi.org/10.1016/j.jvb.2009.03.006

Leite, W. L., & Stapleton, L. M. (2011). Detecting growth shape misspecifications in latent growth models: an evaluation of fit indexes. *The Journal of Experimental Education*, *79*, 361–381. https://doi.org/10.1080/00220973.2010.509369

Li, G., Hou, G., Xie, G., Yang, D., Jian, H., & Wang, W. (2019). Trajectories of self-rated health of chinese elders: A piecewise growth model analysis. *Frontiers in Psychology*, *10*, 583.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling*, *23*, 354–367. https://doi.org/10.1080/10705511.2015.1057285

Lock, E. F., Kohli, N., & Bose, M. (2018). Detecting multiple random changepoints in Bayesian piecewise growth mixture models. *Psychometrika*, *83*, 733–750.

McNeish, D. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using Mplus. *Journal of Educational and Behavioral Statistics*, *41*, 27–56. https://doi.org/10.3102/1076998615621299

McNeish, D. (2020). Relaxing the proportionality assumption in latent basis models for nonlinear growth. *Structural Equation Modeling*, *27*, 817–824. https://doi.org/10.1080/10705511.2019.1696201

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*, 802–829.

Muthén, B. O. (2010). *Bayesian Analysis in Mplus: A brief introduction* [Tech. Rep.]. http://www.statmodel.com/download/IntroBayesVersion%201.pdf

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Ning, L., & Luo, W. (2017). Specifying turning point in piecewise growth curve models: challenges and solutions. *Frontiers in Applied Mathematics and Statistics*, *3*, 19. https://doi.org/10.3389/fams.2017.00019

Patrick, M. E., & Schulenberg, J. E. (2011). How trajectories of reasons for alcohol use relate to trajectories of binge drinking: National panel data spanning late adolescence to early adulthood. *Developmental Psychology*, *47*, 311–317.

Raudenbush, S. W., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387–401.

Rioux, C., Stickley, Z. L., & Little, T. D. (2021). Solutions for latent growth modeling following COVID-19-related discontinuities in change and disruptions in longitudinal data collection. *International Journal of Behavioral Development*, *45*, 463–473. https://doi.org/10.1177/01650254211031631

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of *P* values in composite null models. *Journal of the American Statistical Association*, *95*, 1143–1156.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37–52. https://doi.org/10.1007/BF02294318

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

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*, 310–334.

Smid, S. C., Depaoli, S., & Van De Schoot, R. (2020). Predicting a distal outcome variable from a latent growth model: ML versus Bayesian estimation. *Structural Equation Modeling*, *27*, 169–191. https://doi.org/10.1080/10705511.2019.1604140

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 131–161. https://doi.org/10.1080/10705511.2019.1577140

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. v d (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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. v d (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

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factor [Paper presentation]. Paper presented at the Annual Spring Meeting of the Psychometric Society, Iowa City.

Stern, H. S., & Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, *19*, 2377–2397. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18 < 2377::AID-SIM576 > 3.0.CO;2-1

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. [Database] https://doi.org/10.1007/BF02291170

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., & Van Loey, N. E. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivariate Behavioral Research*, *53*, 267–291.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., Bürkner, P.-C. (2019). *Rank-normalization, folding, and localization: An improved* $\hat{R}$ *for assessing convergence of MCMC*. https://arxiv.org/pdf/1903.08008.pdf

Winter, S. D. (2021). Advanced methods for detecting specification issues in Bayesian structural equation modeling. *UC Merced*. https://escholarship.org/uc/item/118960m4

Winter, S. D., & Depaoli, S. (2022). Sensitivity of Bayesian model fit indices to the prior specification of latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 667–686. https://doi.org/10.1080/10705511.2022.2032078

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*, 183–201.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. University of California. https://www.proquest.com/openview/b0b7f1f7bc043e85742f70df424743ab/1?pq-origsite=gscholar&cbl=18750&diss=y

Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*, 374–383. https://doi.org/10.1177/0165025407077764

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*, 390–420. https://doi.org/10.1177/0149206313501200