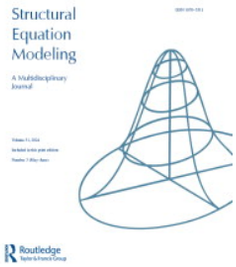

Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

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

Performance of Model Fit and Selection Indices for Bayesian Piecewise Growth Modeling with Missing Data

Ihnwhi Heo, Fan Jia & Sarah Depaoli

To cite this article: Ihnwhi Heo, Fan Jia & Sarah Depaoli (2024) Performance of Model Fit and Selection Indices for Bayesian Piecewise Growth Modeling with Missing Data, Structural Equation Modeling: A Multidisciplinary Journal, 31:3, 455-476, DOI: 10.1080/10705511.2023.2264514

To link to this article: https://doi.org/10.1080/10705511.2023.2264514

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2024, VOL. 31, NO. 3, 455–476
https://doi.org/10.1080/10705511.2023.2264514

Performance of Model Fit and Selection Indices for Bayesian Piecewise Growth Modeling with Missing Data

Ihnwhi Heo, Fan Jia and Sarah Depaoli

University of California, Merced

ABSTRACT

The Bayesian piecewise growth model (PGM) is a useful class of models for analyzing nonlinear change processes that consist of distinct growth phases. In applications of Bayesian PGMs, it is important to accurately capture growth trajectories and carefully consider knot placements. The presence of missing data is another challenge researchers commonly encounter. To address these issues, one could use model fit and selection indices to detect misspecified Bayesian PGMs, and should give care to the potential impact of missing data on model evaluation. Here we conducted a simulation study to examine the impact of model misspecification and missing data on the performance of Bayesian model fit and selection indices (PPP-value, BCFI, BTLI, BRMSEA, BIC, and DIC), with an additional focus on prior sensitivity. Results indicated that (a) increasing the degree of model misspecification and amount of missing data aggravated the performance of indices in detecting misfit, and (b) different prior specifications had negligible impact on model assessment. We provide practical guidelines for researchers to facilitate effective implementation of Bayesian PGMs.

KEYWORDS

Bayesian piecewise growth; missing data; model fit; model misspecification; model selection

The question of how social and behavioral phenomena change over time is at the core of latent growth modeling. The latent piecewise growth model (PGM) is a special case of the latent growth model to describe nonlinear dynamic change processes. It links developmental stages by introducing knots (i.e., connecting points). The PGM is appealing to researchers aiming to model nonlinear trajectories consisting of multiple growth phases. This has led to a proliferation of the PGM in substantive research (e.g., Chung et al., 2017; Hu et al., 2020; Jaggars & Xu, 2016; Kroese et al., 2013; Patrick & Schulenberg, 2011). In the meantime, there have been numerous studies concerning difficulties in estimating nonlinear trends (e.g., Diallo et al., 2014; Grimm et al., 2011), and the PGM is not an exception to the problem (Kohli et al., 2015; Kohli & Harring, 2013). Bayesian methods can be alternatively suited for estimating nonlinear growth models (Lu et al., 2011; Serang et al., 2015; Smid et al., 2020; Zhang et al., 2007), including the PGM (e.g., Kohli et al., 2015; Lock et al., 2018).

The pertinent usage of Bayesian PGMs is closely linked to correct knot placement. Locations of knots should indicate substantively important moments and satisfy theoretical justifications (Flora, 2008; Kwok et al., 2010; Marcoulides, 2018; Ning & Luo, 2017). Knots specified at incorrect locations are tied to the suboptimal representation of true growth patterns and may lead to misleading research conclusions. A major methodological task is hence to assess the fit between a hypothesized Bayesian PGM and data. To accomplish this task of model evaluation, various Bayesian model fit and selection indices (e.g., Asparouhov & Muthén, 2021; Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018) can be utilized to detect model misspecification (Depaoli et al., 2023). In addition, the ubiquity of missing data in longitudinal studies applies to piecewise growth modeling (e.g., Hu et al., 2020; Lee & Rojewski, 2009; Li et al., 2001). In the Bayesian estimation framework, missing data are handled with data augmentation: Observed data are augmented with missing data and then combined with prior distributions for posterior inference (Lee, 2007; Tanner & Wong, 1987). A detailed exposition is provided by Daniels and Hogan (2008) and Gelman et al. (2013) of how data augmentation works for obtaining posterior estimates—under both ignorable missingness (i.e., missing at random [MAR] and missing completely at random [MCAR]) and nonignorable missingness (i.e., missing not at random [MNAR]).

We highlight that a noteworthy aspect is an interplay between the two issues: model misspecification and missing data. Both issues can complicate the process of evaluating Bayesian PGMs. It is known from the growth modeling literature that failure to detect model misfit in the presence of missing data can result in biased growth parameter estimation that misleads conclusions, model evaluation that favors misspecified latent growth models, or both (Lu & Zhang, 2021; Winter & Depaoli, 2022b). Another important element in Bayesian model evaluation is the specification of prior distributions (Cain & Zhang, 2019; Depaoli et al., 2023; Edwards & Konold, 2023; Winter & Depaoli, 2022b). Put

CONTACT Ihnwhi Heo ihnwhi.heo@gmail.com Department of Psychological Sciences, University of California, 5200 N. Lake Road, Merced, CA 95343, USA.
Supplemental data for this article can be accessed online at https://osf.io/agn4x/.

differently, how prior information is incorporated into evaluating Bayesian PGMs may render different consequences in model selection. Thus far, Depaoli et al. (2023) is one notable work that examined the abilities of Bayesian model assessment indices to detect misfit in Bayesian PGMs. Depaoli et al. (2023) revealed that different prior specifications had negligible effects on the overall performance of the indices considered. In addition, their investigation provided valuable insights into the consequences of slope change rates, misspecified growth trajectories, misspecified knot locations, and sample sizes in evaluating Bayesian PGMs. However, their study only utilized complete datasets, and the potential impact of missing data on the evaluation of Bayesian PGMs remains unexplored to the best of our knowledge.

1.1. Novel Contributions

While Depaoli et al. (2023) found little impact of prior specifications on the evaluation of Bayesian PGMs, previous studies highlighted different outcomes of Bayesian model evaluation under varying conditions of prior specifications in the presence of missing data. For instance, Winter and Depaoli (2022b) demonstrated the role of prior distributions in evaluating quadratic latent growth models with missing values. Their findings indicated that, with an increasing amount of missing data, some model fit indices were better at reliably detecting model misspecification when informative priors were correctly placed. Because PGMs are another type of nonlinear latent growth models, it is reasonable to anticipate different consequences when evaluating Bayesian PGMs in the presence of missing data despite the little effect of priors under complete data scenarios found in Depaoli et al. (2023). As evidenced by Winter and Depaoli (2022b), it could be the case that the use of informative and accurate priors helps detect true models compared to weakly informative or diffuse prior distributions. Therefore, our examination takes prior distributions into account in detecting misfit in Bayesian PGMs under missing data scenarios, which is one contribution that sets our study apart.

The usual approach for evaluating the impact of missing data on model evaluation within longitudinal studies has typically involved manipulating the amount and spread of missing data. For instance, Winter and Depaoli (2022b) considered scenarios with 0%, 15%, and 50% as the overall amount of missing data distributed across either a single time point or four consecutive time points. As another example, Shi et al. (2021) included 0%, 15%, and 30% as the total amount of missing data and imposed missingness on the second half of repeated measurements. In both examples, only a single attrition scenario was assumed. However, various forms of attrition are possible in longitudinal studies. To illustrate, with the same 30% of attrition at the end of a study, one could observe a majority of missing data in the early stages of data collection; alternatively, most missing data could occur at later measurement occasions. It is also possible that the proportion of attrition increases at a constant rate, resulting in an aggregate of 30% of missing data at the last time point. Therefore, we referred to Ortega-Azurduy et al. (2008) and considered linear and quadratic functions in missing data generation to reflect various attrition scenarios in practice. Because no methodological studies have yet focused on these different possible forms of longitudinal missing data patterns in the context of Bayesian model evaluation, we consider this aspect as one important factor in the evaluation of Bayesian PGMs, being another novel contribution. In particular, we expect to observe different performances of Bayesian model evaluation tools under different attrition patterns.

In the analysis of piecewise growth patterns, researchers can consider PGMs with either a single knot location (e.g., Hu et al., 2020; Jaggars & Xu, 2016; Li et al., 2019) or multiple knot locations (e.g., Chung et al., 2017; Kroese et al., 2013) to describe separate developmental stages. While our primary focus centers on single-knot PGMs, we also recognize the importance of evaluating Bayesian PGMs with multiple knots. To explore this, we conducted a secondary simulation study to have a preliminary look at complex scenarios and thus contribute to the Bayesian evaluation of PGMs.

1.2. Goals and Organization

Given the utility of the Bayesian PGM, it is imperative that researchers comprehend how model misspecification and missing data impact detecting model misfit in Bayesian PGMs. In addition, the application of the Bayesian PGM is expected to rise due to recent implementations of Bayesian model fit and selection indices into statistical software (Asparouhov & Muthén, 2021; Garnier-Villarreal & Jorgensen, 2020) and openly available code for piecewise modeling (Kohli et al., 2015, 2019). We are thus going to provide a holistic understanding of the performance of model fit and selection indices for Bayesian piecewise growth modeling with missing data by achieving the following two aims: We investigate (a) the performance of Bayesian model fit and selection indices in detecting misspecified Bayesian PGMs in the presence of model misspecification and missing data, and (b) the impact of different prior specifications on the overall performance of Bayesian model fit and selection indices. We assume ignorable missingness (i.e., MAR) in this study, as MAR is a basic assumption of modern missing data techniques (van Buuren, 2018) and it has been commonly studied in the growth curve modeling literature (e.g., Shi et al., 2021; Winter & Depaoli, 2022b).

We outline this article as follows. We first introduce the Bayesian PGM. We then present the Bayesian model fit and selection indices. In the following two sections, we peruse extant literature on the impact of model misspecification and missing data relevant to the current investigation. Next, we describe the design of the simulation study, which is followed by simulation results. The secondary simulation study is then presented. This article ends by summarizing the key findings surrounding the Bayesian model fit and selection procedures, providing guidelines for researchers interested

in implementing Bayesian PGMs, and discussing future research directions.

2. Bayesian Piecewise Growth Modeling

Latent growth models have been extended to represent nonlinear trends (e.g., Blozis, 2007; Grimm et al., 2011; Kohli & Harring, 2013; Ram & Grimm, 2007). However, Diallo et al. (2014) pointed out that nonlinear growth models may face estimation issues such as high rates of non-convergence or inadmissible solutions, particularly when sample sizes are small. The Bayesian approach can reduce problems associated with non-converged and inadmissible estimates by incorporating prior information (Can et al., 2015; Kohli et al., 2015). In terms of piecewise extensions, Bayesian estimation can be even more useful because the specification of knots introduces additional model complexity, which may also affect model estimation (see, e.g., Kohli et al., 2015; Lock et al., 2018). In what follows, we describe the Bayesian approach to PGMs.

2.1. Likelihood Model

Let $\boldsymbol{y}_i$ denote a $J \times 1$ vector of repeatedly measured outcomes for individual i and $\boldsymbol{\eta}_i$ be an $M \times 1$ vector of growth factors for individual i. A PGM can be generally expressed with matrix notation:

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i \text{ with } \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{y}}), \\ \boldsymbol{\eta}_i &= \boldsymbol{\alpha} + \boldsymbol{\zeta}_i \text{ with } \boldsymbol{\zeta}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_{\boldsymbol{\eta}}), \end{aligned} \quad (1)$$

where $\boldsymbol{\Lambda}$ is a $J \times M$ matrix of factor loadings that describe growth rates; $\boldsymbol{\epsilon}_i$ is a $J \times 1$ vector of error terms; $\boldsymbol{\Omega}_{\boldsymbol{y}}$ is a J-dimensional covariance matrix; $\boldsymbol{\alpha}$ is an $M \times 1$ vector of growth factor means; $\boldsymbol{\zeta}_i$ is an $M \times 1$ vector of variance components that reflect the interindividual variation from $\boldsymbol{\alpha}$; and $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$ is an M-dimensional covariance matrix.

An important element in PGMs is the knot location, which can be either fixed or freely estimated. The $\boldsymbol{\Lambda}$ matrix can be used to specify piecewise trajectories, as well as knot locations. To demonstrate, we use a simple PGM in which two linear growth trajectories join at a single knot at occasion k. If we adjust $\boldsymbol{y}_i$ in Equation (1) in scalar terms (e.g., Depaoli et al., 2023; Grimm et al., 2016), we have:

$$y_{ij} = \eta_{0i} + \eta_{1i} \times \min(t_j, k) + \eta_{2i} \times \max(t_j - k, 0) + \epsilon_{ij}. \quad (2)$$

In Equation (2), y_{ij} denotes the outcome of interest for individual i measured at occasion j with $1 \leq j \leq J$; t_j refers to the time metric associated with the occasion j; η_{0i} is the latent intercept; η_{1i} is the first latent linear slope; η_{2i} is the second latent linear slope; and ϵ_{ij} is the error term. The means of the growth factors are denoted by $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)^T$, where respective subscripts correspond to growth factors (e.g., α_1 is the mean of η_{1i}). Next, we can express linear-linear piecewise growth trajectories by specifying the $\boldsymbol{\Lambda}$ matrix such that

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & \min(t_j, k) & \max(t_j - k, 0) \end{bmatrix},$$

where the dimension of the $\boldsymbol{\Lambda}$ matrix is $J \times 3$; the first column defines the intercept, and the second and third columns respectively define the first and second linear slopes that change with time.

For the present investigation, we focus on and account for the PGM in its simplest form in Equation (2): Two linear growth trajectories joined at a single knot. Using this form of PGM with fixed knot location ensured us to examine the performance of Bayesian model fit and selection indices when knot locations could be misspecified, which will be detailed later in the section devoted to the simulation design. It is possible, however, to construct PGMs in more complex formats, for instance, by introducing multiple phases of development followed by the inclusion of more knots (Flora, 2008) or by including nonlinear growth phases (Harring et al., 2021). Corresponding locations of knots can be set *a priori*, freely estimated, or allowed to vary across individuals to describe individual differences (Harring et al., 2021). In addition, Cudeck and Codd (2012) formulated PGMs with disjointed knots, opening up potential applications if growth segments are discontinuous.

2.2. Prior Distribution

Specifying prior distributions is an important task for Bayesian estimation. Our prior settings are defined within the M*plus* software (Muthén & Muthén, 2017), which was the main software used in this paper. There are three types of parameters that need prior specifications. First, the mean of each latent growth factor in $\boldsymbol{\alpha}$ has its associated normal prior distribution:

$$\alpha \sim \mathcal{N}(\mu, \sigma^2),$$

where μ is the mean hyperparameter that determines the location of the prior, and σ^2 is the variance hyperparameter that determines the degree of informativeness. Second, the covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{\eta}}$ between growth factors η_{0i}, η_{1i}, and η_{2i} has its inverse Wishart prior distribution:

$$\boldsymbol{\Omega}_{\boldsymbol{\eta}} \sim \mathcal{IW}(\boldsymbol{\Psi}, \nu),$$

where $\boldsymbol{\Psi}$ is a positive definite matrix; and ν refers to the degrees of freedom. Third, the variances of error terms in $\boldsymbol{\Omega}_{\boldsymbol{y}}$ have corresponding inverse gamma prior distributions:

$$\sigma^2_{\epsilon_{ij}} \sim \mathcal{IG}(a, b),$$

where a is the shape hyperparameter; and b the scale hyperparameter. We present specific prior settings (i.e., hyperparameter settings) in the section devoted to the simulation study.

3. Bayesian Model Fit and Selection Indices

The current section surveys model fit and selection indices for evaluating Bayesian PGMs. More indices exist; however, the indices introduced here are widely studied in prior studies (e.g., Cain & Zhang, 2019; Depaoli et al., 2023; Edwards & Konold, 2023; Winter & Depaoli, 2022a, 2022b). Readers are referred to references therein at each subsection for more technical details.

3.1. Information Criteria (BIC and DIC)

Information criteria are model selection indices to select the best model that represents observed data among multiple competing models. We introduce two information criteria: the Bayesian information criterion (BIC; Schwarz, 1978) and the deviance information criterion (DIC; Spiegelhalter et al., 2002). These two selection indices are used to compare either nested or non-nested models, and the model with the smallest BIC or DIC value is preferred over the others.

Let $\boldsymbol{\theta}$ denote a vector of parameters, and let $\boldsymbol{y}$ replace $\boldsymbol{y}_i$ for simplicity. The BIC is the approximation of the Bayes factor (Kass & Raftery, 1995) and is formulated as follows:

$$\text{BIC} = -2\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{ML}})\} + k\log n,$$

where $\hat{\boldsymbol{\theta}}_{\text{ML}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$; k is the number of parameters in the model; n is the sample size; $\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{ML}})\}$ is the log likelihood based on the maximum likelihood estimates of $\boldsymbol{\theta}$; and $k\log n$ is the model complexity term that penalizes models with more number of parameters.[1]

The DIC is defined in the following way:

$$\text{DIC} = -2\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{EAP}})\} + 2p_D,$$

where $\hat{\boldsymbol{\theta}}_{\text{EAP}}$ is the posterior mean of $\boldsymbol{\theta}$, used to define the term $\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{EAP}})\}$ that evaluates the log likelihood; and p_D is an estimate of the effective number of parameters. This latter term, p_D, is the model complexity term that is computed in the following formulation:

$$p_D = 2(\log\{p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{EAP}})\}) - E_{p(\boldsymbol{\theta}|\boldsymbol{y})}\left[\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}\right],$$

where, on the right side, the second term $E_{p(\boldsymbol{\theta}|\boldsymbol{y})}[\log\{p(\boldsymbol{y}|\boldsymbol{\theta})\}]$ refers to the average log likelihood over the posterior distribution of $\boldsymbol{\theta}$.

Within the Bayesian estimation framework, prior distributions should be specified by adjusting the hyperparameter values. These values can be determined either through fixed numbers or distributional forms. Using distributions to probabilistically sample hyperparameter values introduces additional parameters for the specification of such distributions. This process can complicate the Bayesian analyses by inflating the number of parameters for any given model. A key distinction between the BIC and DIC lies in their treatment of counting those parameters. The BIC considers all the parameters either for the model or for priors, which can over-penalize the model complexity. In contrast, the DIC can be formulated to include only the effective number of parameters, as determined by researchers. Therefore, the DIC has advantages in handling a large number of parameters (see also Asparouhov et al., 2015).

[1]We provide two cautionary notes for our readers. First, when we use the term "Bayesian model selection indices", we are referring to both the BIC and DIC, not just the BIC alone. Second, the computation of the BIC relies on the maximum likelihood estimates of the parameters. That said, the BIC cannot be considered a purely "Bayesian measure" given that its specification is not based on the posterior distribution (Gelman et al., 2014). However, we included the BIC in the current investigation because (1) it is provided as default in M*plus* when using the Bayesian estimation framework, and (2) its performance was investigated in previous studies on Bayesian model evaluation (Depaoli et al., 2023; Winter & Depaoli, 2022a, 2022b).

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

If a model fits observed data well, future observations predicted from the model should be similar to the observed data. The PPP-value is the model fit index that is based on such idea; hence, we simulate future data from the distribution called the posterior predictive distribution and compare the future data to the observed data. The posterior predictive distribution is the probability density that future observations are expected to arise from and can be written as follows:

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

where $\boldsymbol{y}^{\text{rep}}$ denotes future data or replicated data; and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution.

The procedure of sampling from the posterior predictive distribution is based on the MCMC methods. At each iteration s, sampled $\boldsymbol{\theta}^{(s)}$ is imputed in $p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{\theta})$ to simulate each replicated dataset of the same size as the observed data, and this sampling scheme is performed iteratively. The resulting multiple replicated datasets should be compared to the observed data to test if each pair of replicated data and observed data has resemblance to each other. A discrepancy function—a function of data and parameters—is used to evaluate and compare model fit to replicated versus observed data between the estimated model $\mathcal{M}_0$ and the unconstrained model $\mathcal{M}_1$ (Muthén, 2010). The most commonly used discrepancy function is the likelihood ratio test (LRT) chi-square, which has been formulated in M*plus* as follows [2]:

$$D(\cdot) = D(\boldsymbol{y}, \mu_1, \boldsymbol{\Sigma}_1, \mu_0, \boldsymbol{\Sigma}_0) = \mathcal{L}(\boldsymbol{y}|\mu_1, \boldsymbol{\Sigma}_1) - \mathcal{L}(\boldsymbol{y}|\mu_0, \boldsymbol{\Sigma}_0), \tag{3}$$

where the subscripts 1 and 0 of μ and $\boldsymbol{\Sigma}$ respectively correspond to $\mathcal{M}_1$ (i.e., unconstrained model) and $\mathcal{M}_0$ (i.e., estimated model); $\mathcal{L}(\boldsymbol{y}|\mu_1, \boldsymbol{\Sigma}_1)$ is the log likelihood of $\boldsymbol{y}$ based on the multivariate normal distribution with mean μ_1 and the covariance matrix $\boldsymbol{\Sigma}_1$; and $\mathcal{L}(\boldsymbol{y}|\mu_0, \boldsymbol{\Sigma}_0)$ refers to the log likelihood defined in the same way using μ_0 and $\boldsymbol{\Sigma}_0$. The discrepancy function in Equation (3) for the observed data is then expressed as:

$$D(\cdot)_s^{\text{obs}} = D_s^{\text{obs}}(\boldsymbol{y}^{\text{obs}}, \mu_{1,s}(\boldsymbol{y}^{\text{obs}}), \boldsymbol{\Sigma}_{1,s}(\boldsymbol{y}^{\text{obs}}), \mu_{0,s}, \boldsymbol{\Sigma}_{0,s}),$$

where $\mu_{1,s}(\boldsymbol{y}^{\text{obs}})$ and $\boldsymbol{\Sigma}_{1,s}(\boldsymbol{y}^{\text{obs}})$ refer to a random draw of the $\mathcal{M}_1$ parameter estimates for $\boldsymbol{y}^{\text{obs}}$ at iteration s; and $\mu_{0,s}$ and $\boldsymbol{\Sigma}_{0,s}$ refer to the $\mathcal{M}_0$-implied mean and covariance matrix obtained from the $\mathcal{M}_0$ at iteration s. In a similar vein, the discrepancy function for the replicated data is:

$$D(\cdot)_s^{\text{rep}} = D_s^{\text{rep}}(\boldsymbol{y}^{\text{rep}}, \mu_{1,s}(\boldsymbol{y}^{\text{rep}}), \boldsymbol{\Sigma}_{1,s}(\boldsymbol{y}^{\text{rep}}), \mu_{0,s}, \boldsymbol{\Sigma}_{0,s}),$$

where $\mu_{1,s}(\boldsymbol{y}^{\text{rep}})$ and $\boldsymbol{\Sigma}_{1,s}(\boldsymbol{y}^{\text{rep}})$ indicate a random draw of the $\mathcal{M}_1$ parameter estimates for $\boldsymbol{y}^{\text{rep}}$ at iteration s. Both

[2]As of M*plus* version 8.4, missing values remain missing when computing the discrepancy function (see Asparouhov & Muthén, 2021). The replicated data generated at each iteration using the incomplete observed data mimics the missing data patterns. The same missing data patterns between the observed and replicated data ensure the comparability of both data under the null hypothesis that $\mathcal{M}_0$ is true. In addition, Asparouhov and Muthén (2021) pointed out that replicated data can be considered MAR under the assumption that the actual data is MAR. Therefore, comparability of the discrepancy function for the observed and replicated data is ensured.

discrepancy function terms—$D(\cdot)_s^{\text{obs}}$ and $D(\cdot)_s^{\text{rep}}$—are used to calculate the PPP-value:

$$\text{PPP-value} = p(D(\cdot)_s^{\text{rep}} > D(\cdot)_s^{\text{obs}}|\boldsymbol{y}) \approx \frac{1}{S}\sum_{s=1}^{S}\delta_s, \quad (4)$$

where S is the number of iterations of the chain; and δ_s is realized based on the following condition:

$$\delta_s = \begin{cases} 1 & \text{if } D(\cdot)_s^{\text{rep}} > D(\cdot)_s^{\text{obs}}, \\ 0 & \text{otherwise.} \end{cases}$$

The PPP-value in Equation (4) is interpreted as the proportion of replicated datasets whose discrepancy statistics exceed discrepancy statistics applied to the observed data. PPP-values of 0.5 indicates a good model fit to the data, meaning 50% of replicated datasets have discrepancy statistics greater than those of the observed data. Extremely low PPP-values close to 0 indicate model misspecification, signaling most of replicated datasets have greater discrepancy statistics compared to those of the observe data. A common practice is to use .05 as a cutoff criterion such that PPP-values >.05 indicate adequate model fit (Asparouhov & Muthén, 2010).

3.3. Bayesian Approximate Fit Indices

Three approximate fit indices—the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI)—have been recently implemented within the Bayesian scheme (Asparouhov & Muthén, 2021; Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018). These indices depend on the calculation of the PPP-value and are computed at each iteration s of the Markov chain, resulting in the posterior distributions of each fit index.

Bayesian RMSEA (BRMSEA) evaluates the badness-of-fit such that it measures the extent to which a hypothesized model deviates from the perfect model in capturing observed data; hence is also referred to as the absolute fit index. BRMSEA is calculated as follows:

$$\text{BRMSEA}_s = \sqrt{\max(0, \frac{D(\cdot)_s^{\text{obs}} - p^*}{(p^* - p_D)n})},$$

where $D(\cdot)_s^{\text{obs}}$ is the discrepancy function for the observed data at iteration s; and p^* is the number of nonredundant sample moments. BRMSEA captures model misfit as the rescaled discrepancy at iteration s and produces realized values that are chi-square distributed (see Garnier-Villarreal & Jorgensen, 2020).

Bayesian CFI (BCFI) and Bayesian TLI (BTLI) are incremental fit indices such that they evaluate the relative improvement of the fit of the target model along a continuum between the baseline model (i.e., all covariances are fixed to 0) and the saturated model. Indices that take on values near 0 indicate that the target model does not fit the data well compared to the baseline model; indices with values closer to 1 indicate that the target model shows good fit to the data compared to the baseline model. BCFI is formulated as follows:

$$\text{BCFI}_s = 1 - \frac{D(\cdot)_{T,s}^{\text{obs}} - p^*}{D(\cdot)_{B,s}^{\text{obs}} - p^*},$$

where $D(\cdot)_{T,s}^{\text{obs}}$ refers to the target model discrepancy function for the observed data at iteration s; and $D(\cdot)_{B,s}^{\text{obs}}$ refers to the baseline model discrepancy function for the observed data at iteration s. BTLI, on the other hand, is computed in the following way:

$$\text{BTLI}_s = \frac{\frac{D(\cdot)_{B,s}^{\text{obs}} - p_{D_B}}{p^* - p_{D_B}} - \frac{D(\cdot)_{T,s}^{\text{obs}} - p_{D_T}}{p^* - p_{D_T}}}{\frac{D(\cdot)_{B,s}^{\text{obs}} - p_{D_B}}{p^* - p_{D_B}} - 1},$$

where p_{D_B} is the model complexity term defined for the baseline model; and p_{D_T} is the model complexity term defined for the target model.

4. The Impact of Model Misspecification

Within the framework of piecewise growth modeling, a model can be misspecified in either the mean structure or the covariance structure (Depaoli et al., 2023; Leite & Stapleton, 2011; Ning & Luo, 2017). While multiple sources of misspecification are possible within either structure, we particularly focus on the mean structure, which can be teased into misspecification in growth trajectories (latent means) and knot locations in more detail to look into misspecification scenarios in PGMs. An example of a misspecified growth trajectory is to fit a linear growth model when the true model reflects piecewise growth patterns. A misspecified knot location arises when knot locations are specified to time points that are different from true changepoints. These sources of misspecification impact how model fit and selection indices perform to detect model misfit.

Misspecification in the mean and covariance structures mostly received attention in the literature on latent growth modeling. A general finding is that frequentist RMSEA, CFI, and TLI were more sensitive to the misspecified mean structure compared to the misspecified covariance structure (Wu et al., 2009; Wu & West, 2010), and both sources of misspecification are intertwined with each other (Wu et al., 2009). Between the three frequentist indices, RMSEA was better at detecting model misfit compared to CFI and TLI (Yu, 2002). These patterns were similarly observed in Bayesian literature in that BCFI and BTLI were not sensitive to detecting model misspecification (Winter & Depaoli, 2022b). Compared to these two Bayesian approximate fit measures, the PPP-value could detect misspecified mean and covariance structures (Fay et al., 2022; Winter & Depaoli, 2022b). Winter and Depaoli (2022b) additionally observed that DIC was more likely to favor the correctly specified models than BIC.

Because PGMs are extensions of latent growth models, comparable observations have been found on the performance of model assessment measures. For instance, Leite and Stapleton (2011) examined the performance of frequentist approximate fit indices when a PGM was misspecified in

the mean and the covariance structures. Whereas CFI and TLI were more sensitive to the misspecified mean structure than the misspecified covariance structure, both indices were insensitive to the sample size or the severity of misspecification and retained most of the misspecified model. RMSEA, however, performed better than CFI and TLI in detecting misspecified models.

Another source of misspecification in PGMs is knot locations, which was examined by Ning and Luo (2017). Their true models were linear-linear PGMs with knot locations specified at time point 3, 3.5, 4, or 5. By fixing the location of knot of the analysis model at time point 3, they had either correctly specified model or misspecification of either 0.5, 1, or 2 time points. They found that frequentist RMSEA, CFI, and TLI did not perform well in detecting misspecified knot locations irrespective of sample sizes and the severity of misspecifications.

The poor performances of approximate fit indices were also found within the Bayesian context. In particular, Depaoli et al. (2023) misspecified knot locations as well as growth trajectories. There were three data-generating models, all of which were linear-linear PGMs with their knot locations fixed at time point 4. The only difference between the three models was the magnitude of the change in the second growth trajectory, which was small, medium, or large based on effect sizes. The analysis model, on the other hand, was either a correctly specified model or three misspecified models such that the true knot location was ignored or misspecified at time points 3 or 5. Consistent with Fay et al. (2022) and Winter and Depaoli (2022b), the PPP-value could reliably detect model misspecifications for larger sample sizes and for a larger change in the second growth phase. However, BCFI, BTLI, and BRMSEA showed unreliable performance in detecting misfit. Between BIC and DIC, BIC could not pinpoint severe misspecifications with ignored knots from the true model for smaller sample sizes, indicating the outperformance of DIC.

It is evident from the literature that model misspecification can influence the performance of various model evaluation tools. We highlight that missing data is another critical factor that may have substantial consequences in detecting model misfit. In the next section, we focus on model evaluation issues surrounding missing data.

5. The Impact of Missing Data

The presence of missing data in longitudinal studies is attributable to many reasons; for example, missingness is planned by study design (Rhemtulla et al., 2014; Wu & Jia, 2021), or participants drop out of a study at the midway point (Nicholson et al., 2017; Twisk & de Vente, 2002). The latter scenario, called attrition, is a typical pattern of longitudinal missing data and is our focus.

Literature on the impact of missing data on model fit and selection indices is scarce, particularly when it comes to the PGM, regardless of being frequentist or Bayesian. In the frequentist latent growth modeling literature, Shi et al. (2021) studied the performance of fit indices in the presence of missing data in a linear latent growth model. They manipulated the percentage of MCAR or MAR data by imposing either 0%, 10%, or 30% of missingness. Their results showed that frequentist RMSEA and CFI could not detect a correct model, although RMSEA was slightly better at detecting model misfit than CFI. Importantly, worsened performance of these indices was observed for smaller samples or increasing amounts of missing data.

The negative impact of increasing amounts of missing data on model assessment measures was similarly observed within the Bayesian latent growth modeling literature. Winter and Depaoli (2022b) focused on a quadratic model and differed the amount of missing data by either 0%, 15%, or 50% based on the MAR assumption. According to their results, the ability of the PPP-value to detect misspecification decreased when the amount of missing data increased. BCFI and BTLI showed a similar pattern as the PPP-value in that the performance of these two Bayesian fit indices was slightly aggravated when the amount of missing data increased. BRMSEA, on the other hand, was relatively not affected by the presence of missing data except for a condition where 50% of missing data existed. In addition, Bayesian model selection indices were under the influence of the amount of missing data. The percentage of incorrectly favoring misspecified models according to the BIC and DIC values increased when the amount of missing data was larger (see Celeux et al., 2006, for explanations of poor performance of DIC in the presence of missing data).

We note the scarcity of research investigating the impact of missing data on the performance of model fit and selection indices in Bayesian PGMs, which reaffirms the importance of our current investigation. Uncovering this unexplored area will contribute to a more complete understanding of the issues related to model evaluation for Bayesian piecewise growth modeling.

6. A Simulation Study

We conducted a Monte Carlo simulation study to evaluate the performance of Bayesian model fit and selection indices in Bayesian PGMs. Specifically, we examined how model misspecification and missing data affect these indices under different prior specifications. We implemented the simulation study with a fully crossed factorial design with four factors: sample sizes (3 levels), missing data (7 levels), knot placement (4 levels), and prior specifications (3 levels), resulting in a total of 252 cells. The details of the simulation design, including the population model, are described in the subsequent paragraphs.

6.1. The Population Model

Our data-generating model was a linear-linear PGM with seven repeated measurements. The location of the knot was at the fourth time point ($k = 3$). Population parameter values were determined following Depaoli et al. (2023). We present the path diagram in Figure 1. The mean of the second slope factor reflects a large change after the first

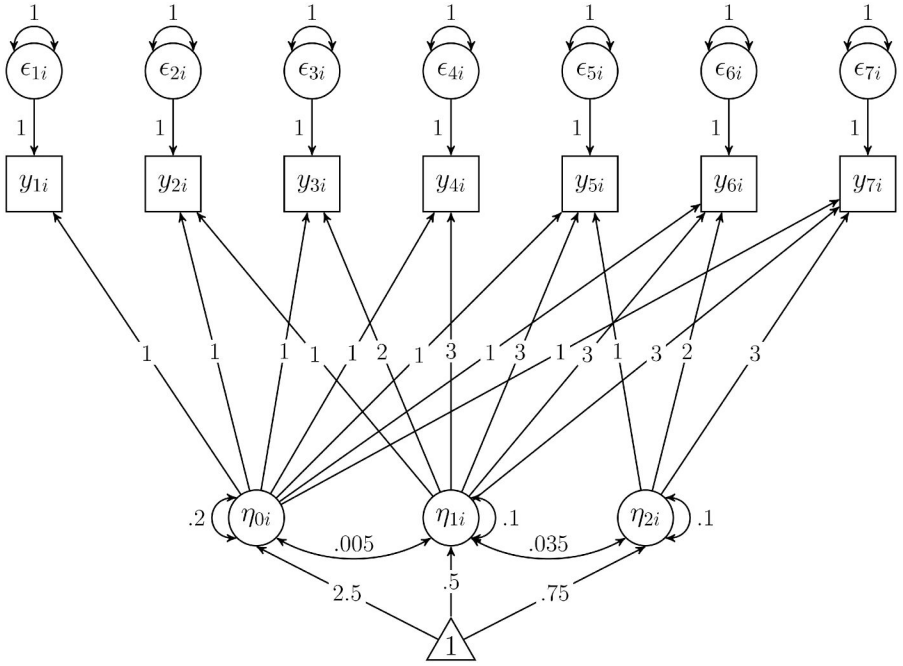

Figure 1. A population model.

slope based on the standardized effect size calculation (Cohen, 2002; Raudenbush & Liu, 2001).

6.2. Sample Size

Sample size is known to be an important factor that affects the performance of Bayesian model fit and selection indices (e.g., Asparouhov & Muthén, 2021; Cain & Zhang, 2019; Depaoli et al., 2023; Garnier-Villarreal & Jorgensen, 2020; Shi et al., 2019; Winter & Depaoli, 2022a, 2022b). We included the following levels of the sample size condition: 50 (small), 150 (medium), and 500 (large). This range of sample sizes is based on Winter and Depaoli (2022b) and covers sample sizes typically encountered in longitudinal studies.

6.3. Missing Data

We considered missing data conditions in terms of the pattern and proportion of missing values. First, the missingness followed attrition patterns: The presence of missing values at a certain time point guaranteed missing values at subsequent time points (Nicholson et al., 2017; Twisk & de Vente, 2002). We modeled the probability of attrition across repeated measurements using either linear or quadratic functions (Ortega-Azurduy et al., 2008). Whereas a linear function represents a constant increase of missing values across repeated measurements, a quadratic function can express either an accelerated increase (i.e., missing values concentrated at the end of the study) or a decelerated increase (i.e., missing values concentrated at the start of the study) of missing values. Second, the proportion of missing values was set at 0% (i.e., no missing values), and 30%, or 70% missingness at the last time point. These proportions were determined based on studies by Gustavson et al. (2012) and Wu et al. (2016) to mirror typical amount of missing data in longitudinal studies.

This ended up with 7 missing data conditions: no missing values (Complete), missing values concentrated at the start of the study with 30% in the end (Concen-S-30), missing values at a constant increase with 30% in the end (Constant-30), missing values concentrated at the end of the study with 30% in the end (Concen-E-30), missing values concentrated at the start of the study with 70% in the end (Concen-S-70), missing values at a constant increase with 70% in the end (Constant-70), and missing values concentrated at the end of the study with 70% in the end (Concen-E-70). Figure 2 describes three levels of the missing data condition where 70% of data are missing at the last time point (i.e., Concen-S-70, Constant-70, and Concen-E-70). Another plot that contains other three levels with 30% missing data at the last time point is available as Online Supplementary Material[3].

We used the following formula to determine the proportion of missing data at each time point ($t_j = 0, 1, \ldots, 6$):

$$\begin{cases} -\frac{5}{6}(t_j-6)^2+30 & \text{if Concen}-\text{S-30}, \\ 5t_j & \text{if Constant-30}, \\ \frac{5}{6}t_j^2 & \text{if Concen}-\text{E-30}, \\ -\frac{35}{18}(t_j-6)^2+70 & \text{if Concen}-\text{S-70}, \\ \frac{35}{3}t_j & \text{if Constant-70}, \\ \frac{35}{18}t_j^2 & \text{if Concen}-\text{E-70}. \end{cases} \quad (5)$$

[3]Online Supplementary Materials are available at the Open Science Framework project (https://osf.io/agn4x/).

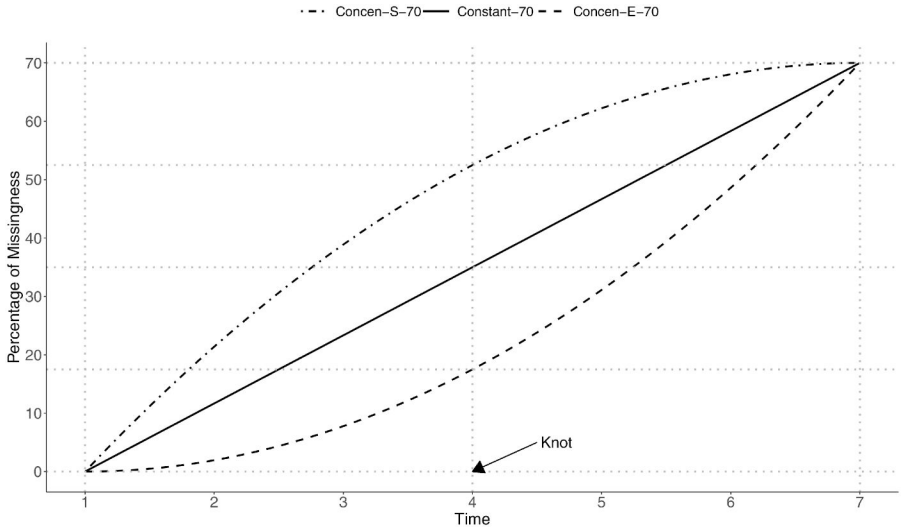

Figure 2. Attrition pattern when 70% of data are missing.

We then used logistic regression (e.g., Agresti, 2012) to determine the missing data indicator ($\mathcal{R}_{ij}$) of individual i at occasion j based on the standardized score at the first measurement occasion to satisfy the MAR assumption:

$$p(\mathcal{R}_{ij} = 1|z_{i1}) = \frac{e^{b_{0,j}+b_1 z_{i1}}}{1 + e^{b_{0,j}+b_1 z_{i1}}}, \quad (6)$$

where z_{i1} refers to a standardized y_{i1}. In Equation (6), the intercept $b_{0,j}$ determines the amount of missing data at occasion j, and the slope b_1 indicates the relationship between the standardized predictor and the probability of missingness. We first chose a slope coefficient b_1 in a way that a squared correlation of 0.40 is produced; this way, we ensured a sufficiently strong relationship between the cause of missingness and the underlying probability for missing data (e.g., Enders & Mansolf, 2018; Winter & Depaoli, 2022b). Next, we determined an intercept $b_{0,j}$ to obtain the desired proportion of missing data at each time point. Using the slope and intercept values determined, we used a logit link to produce a vector of probabilities. These probabilities were used as success probabilities of a binomial distribution to obtain missing data indicators. If the indicators were equal to 1, existing values were removed to treat them as missing values. This process was repeated to reach the desired proportions of missing data across all missing data conditions.

6.4. Knot Placement

We included 4 levels of knot placement to represent either correctly specified or misspecified location of knots. In the analysis model, we placed knot at the true location (i.e., fourth time point), one time point before the true location (i.e., third time point), one time point after the true location (i.e., fifth time point); or we completely ignored knot. Knot placement at the true location reflects correct knot placement, but the other three levels represent misspecified knot placement. Ignoring knot placement means fitting linear growth models instead of PGMs.

6.5. Prior Specification

To understand the prior sensitivity of the performance of Bayesian model fit and selection indices, 3 levels of different prior specifications were included: (1) diffuse prior; (2) informative prior put at the accurate location (informative-accurate prior); and (3) weakly informative prior put at the inaccurate location (weakly informative-inaccurate prior). In all levels, we only differed the priors put on the mean of the latent growth factors; hence, we manipulated values of the mean and variance hyperparameters of the normal prior.

For explanatory purposes, let the subscripts of each hyperparameter refer to the corresponding growth factors; for instance, μ_0 and σ_0^2 respectively refer to the mean and the variance hyperparameters of the normal prior of the intercept. For the diffuse prior, we used the default prior setting in M*plus*: $\mu_0 = \mu_1 = \mu_2 = 0$ and $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 10^{10}$. For the informative-accurate prior, the mean hyperparameters were equal to true population values (i.e., $\mu_0 = 2.5; \mu_1 = 0.5;$ and $\mu_2 = 0.75$), and the variance hyperparameters were 10% of the true population means (i.e., $\sigma_0^2 = 0.25; \sigma_1^2 = 0.05;$ and $\sigma_2^2 = 0.075$). For the weakly informative-inaccurate prior, the mean hyperparameters were shifted upward by adding 3 times the square root of the variance hyperparameters (i.e., $\mu_0 = 4; \mu_1 = 1.171;$ and $\mu_2 = 1.572$), and the variance hyperparameters were 50% of the true population means (i.e., $\sigma_0^2 = 1.25; \sigma_1^2 = 0.25;$ and $\sigma_2^2 = 0.375$). For the other parameters than the mean of the growth factors, the default prior specifications in M*plus* were used.

6.6. Software and Bayesian Estimation

We used the R `lavaan` package (Rosseel, 2012) for data generation with 1,000 replications per cell, and used M*plus* version 8.6 (Muthén & Muthén, 2017) for Bayesian estimation. We implemented the Gibbs sampler to generate two MCMC chains each consisting of 20,000 iterations, with a thinning interval of 1. The first 10,000 iterations in each chain were discarded as the burn-in period of the chain. We

inspected $\hat{R}$ convergence diagnostics to assess chain convergence (Vehtari et al., 2021), with their values <1.05 as a cutoff (e.g., Depaoli et al., 2023).

6.7. Outcomes of Interest

To evaluate the performance of Bayesian model fit and selection indices, we first extracted values of the following Bayesian model fit indices: the PPP-value, BRMSEA, BCFI, and BTLI across all the simulation conditions. We also extracted the 90% credible intervals (CIs) of the three Bayesian approximate fit indices (i.e., BCFI, BTLI, and BRMSEA) to assess their sensitivity to model misspecification and missing data. Depending on where the entire 90% CIs are located based on the cutoff values, we categorized either conclusively good, inconclusive, or conclusively poor model fit. For a conclusively good model fit, the entire 90% CIs is below .06 for BRMSEA and above .95 for BCFI and BTLI; for an inconclusive model fit, the cutoff values are included within the 90% CIs; and for a conclusively poor model fit, the entire 90% CIs fall on the other side of the cutoff values. We expect to observe higher rates of a good model fit when the model was correctly specified and when the presence of missing data has no adverse effects; on the other hand, higher rates of a poor model fit will be obtained if the model fit indices are sensitive to model misspecification and missing data. For Bayesian model selection indices (i.e., BIC and DIC), we calculated the proportion of replications that favored the correctly specified model over misspecified models. The difference in values of both information criteria was also calculated. If these two model selection indices are sensitive to detecting misspecified models and are not affected by missing data, the proportion of favoring the correctly specified model is expected to be higher.

7. Results

The percentage of replications that converged, with $\hat{R}$ diagnostics over all parameters less than 1.05, ranged from 81.4% to 100% (mean = 97.1% and median = 99.5%). The choice of priors did not substantially impact the convergence rates. For each of the diffuse, informative-accurate, and weakly informative-inaccurate priors, the mean percentage of converged replications were 97.2%, 97.4%, and 96.7%, respectively. We report replications with converged results. We also observed very similar results between the BCFI and BTLI; therefore, we only present the BCFI results to save space. Full results including BTLI can be accessed online as Supplementary Material.

7.1. Assessing Model Fit Using Index Values

Well-performing model fit indices should alarm model misfit if models were misspecified. We created boxplots in Figures 3–5 to evaluate the performance of Bayesian model fit indices based on index values. These figures have rows that represent sample sizes and columns that correspond to each level of the missing data condition. In addition, the x-axis represents prior specifications, and the y-axis indicates values of the model fit indices. Four grouped boxes within each prior specification condition correspond to either correctly specified or misspecified models.

7.1.1. PPP-Value

Figure 3 shows results for the PPP-value. There are two lines in each boxplot. The solid horizontal line represents the PPP-value of 0.5 that indicates a good fit; the dashed horizontal line, on the other hand, represents the PPP-value of 0.05, which is a usual cutoff for a poor model fit.[4] The PPP-value performs well when the PPP-values of correctly specified models hover around 0.5, and those of the other misspecified models are centered around or below 0.05.

We first focus on conditions when there were no missing values (i.e., the first column). When $n = 50$, the PPP-value could not detect model misspecification because every model fitted the data well. When the sample size increased from 50 to 150, the PPP-values from the correct model specification hovered around 0.5, meaning the PPP-value could detect correctly specified models. For misspecified models, the PPP-value could reliably detect the severely misspecified models (i.e., ignored knot), but it could not detect the misspecification in terms of knot misplacement. When $n = 500$, the PPP-value performed well in detecting model misspecification in that the PPP-values of the correctly specified model were centered at 0.5, and those of the other misspecified models were close to or even lower than 0.05. Under this condition, all forms of model misspecification could be correctly detected. When there were 30% of missing values (i.e., columns 2–4), we observed that the performance of the PPP-value worsened, even for the largest sample size condition: For $n = 500$, the PPP-values became unreliable because those of misplaced knots approached 0.5 whereas they could still detect severe misspecification. When the amount of missing data was 70% (i.e., the fifth through the seventh column), the performance of the PPP-value was even more aggravated compared to the 30% of missing data conditions. The PPP-value could still detect misspecified models with the ignored knot when $n = 500$, with an increased number of outliers toward 0.5; on the other hand, the PPP-value could not detect misplaced knots. We also found that, when missing values were concentrated at the start of the study, the PPP-values of misspecified models deviated more from 0.05 and shifted toward 0.5. This indicates that having a lot of missing data in the beginning stages of the study can worsen the performance of the PPP-value. Finally, across all the simulation conditions, distributions of the PPP-values were not different across different prior specifications. This pattern reveals that different prior distributions have little impact on the performance of the PPP-value.

[4]We highlight that this value is not a strict cutoff; nor should it be. We are not advocating for implementation of frequentist cutoffs for these Bayesian indices. Instead, the use of cutoff values is rather to facilitate the intuitive interpretation of results that represents the likely way the indices will be used and interpreted in practice. We take this approach in interpreting the other Bayesian model fit indices as well.

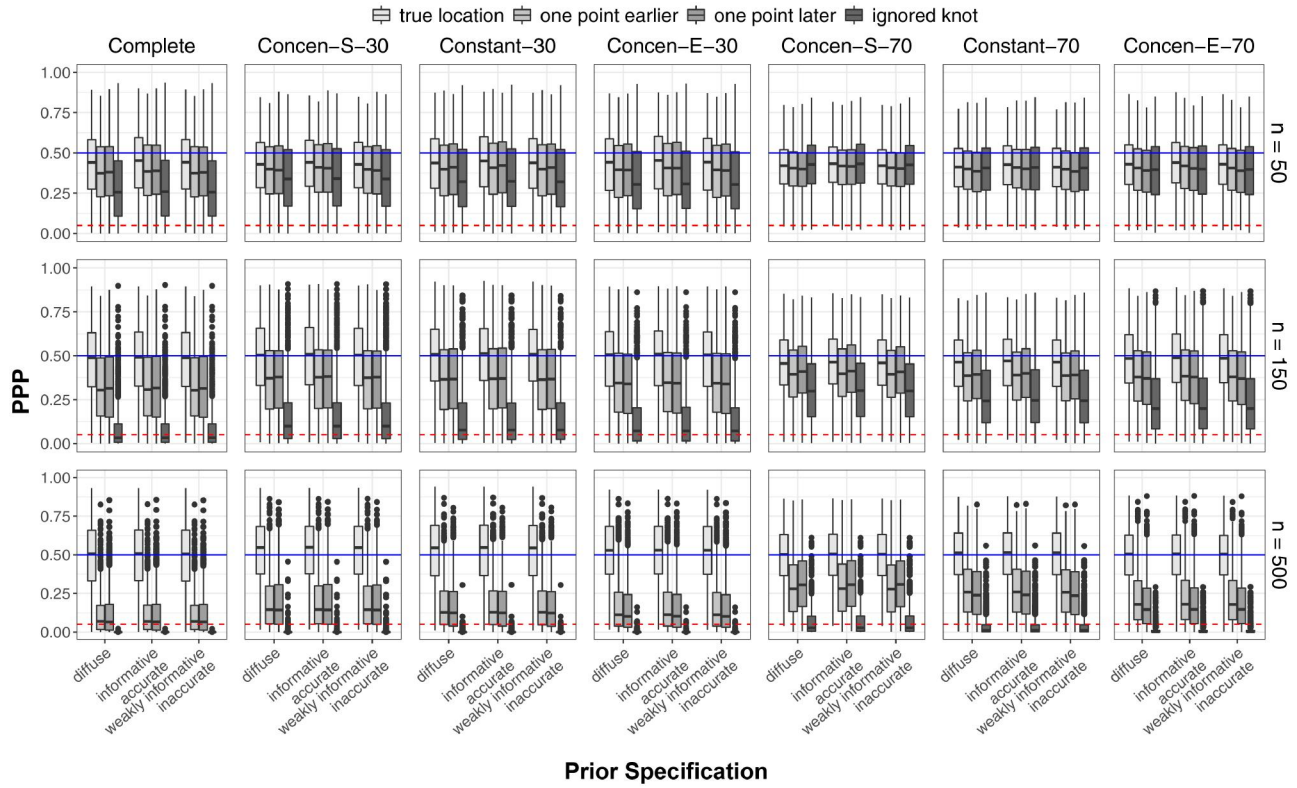

Figure 3. The PPP-value across simulation conditions.

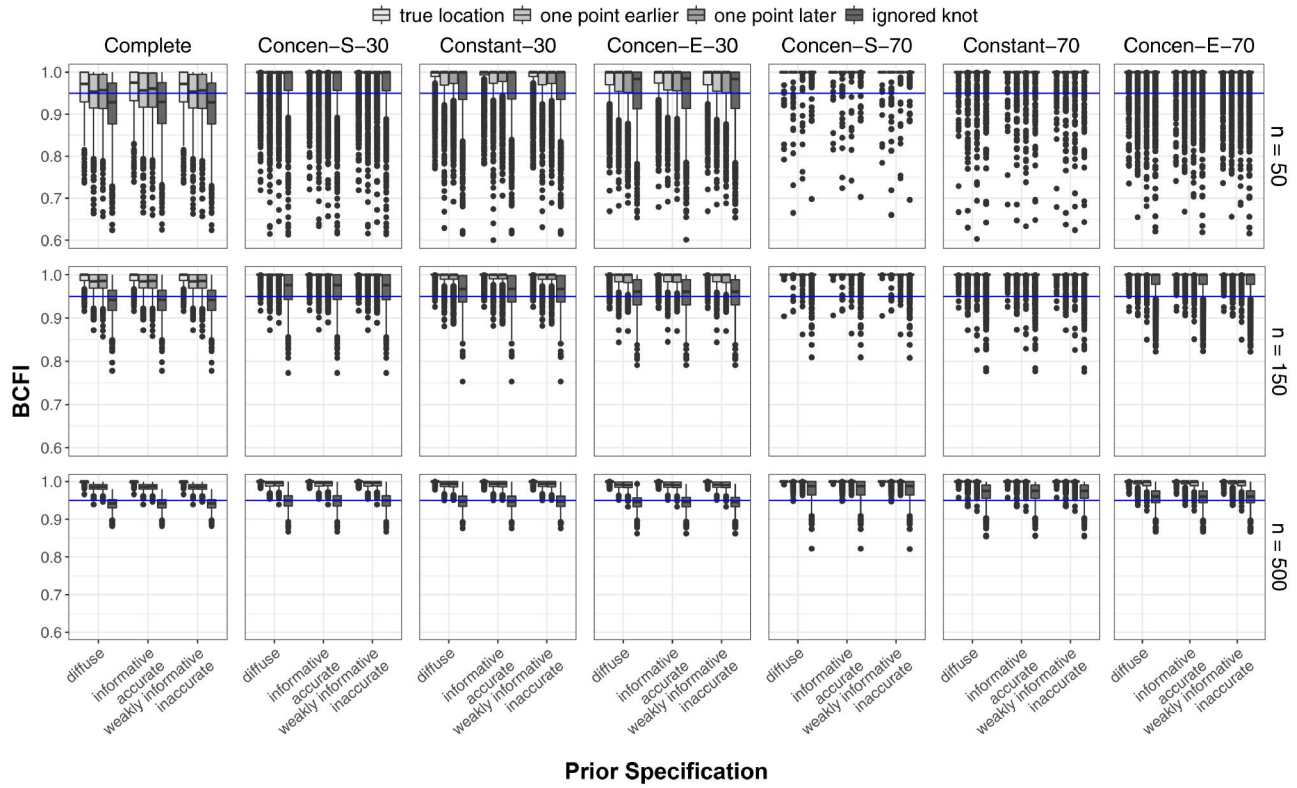

Figure 4. The BCFI across simulation conditions. The range of the y-axis is set from 0.6 to 1 to enhance the interpretability of the results across simulation conditions.

7.1.2. BCFI (and BTLI)

We present results for the BCFI in Figure 4. There is one solid horizontal line at 0.95 in all boxplots, and this is a cut-off to determine a good model fit for the BCFI and BTLI. If values of the BCFI and BTLI $\geq$0.95, a model shows a good fit to the data. We additionally note that we modified the range of the *y*-axis from 0.6 to 1 to enhance the interpretability of the findings. The BCFI and BTLI are considered

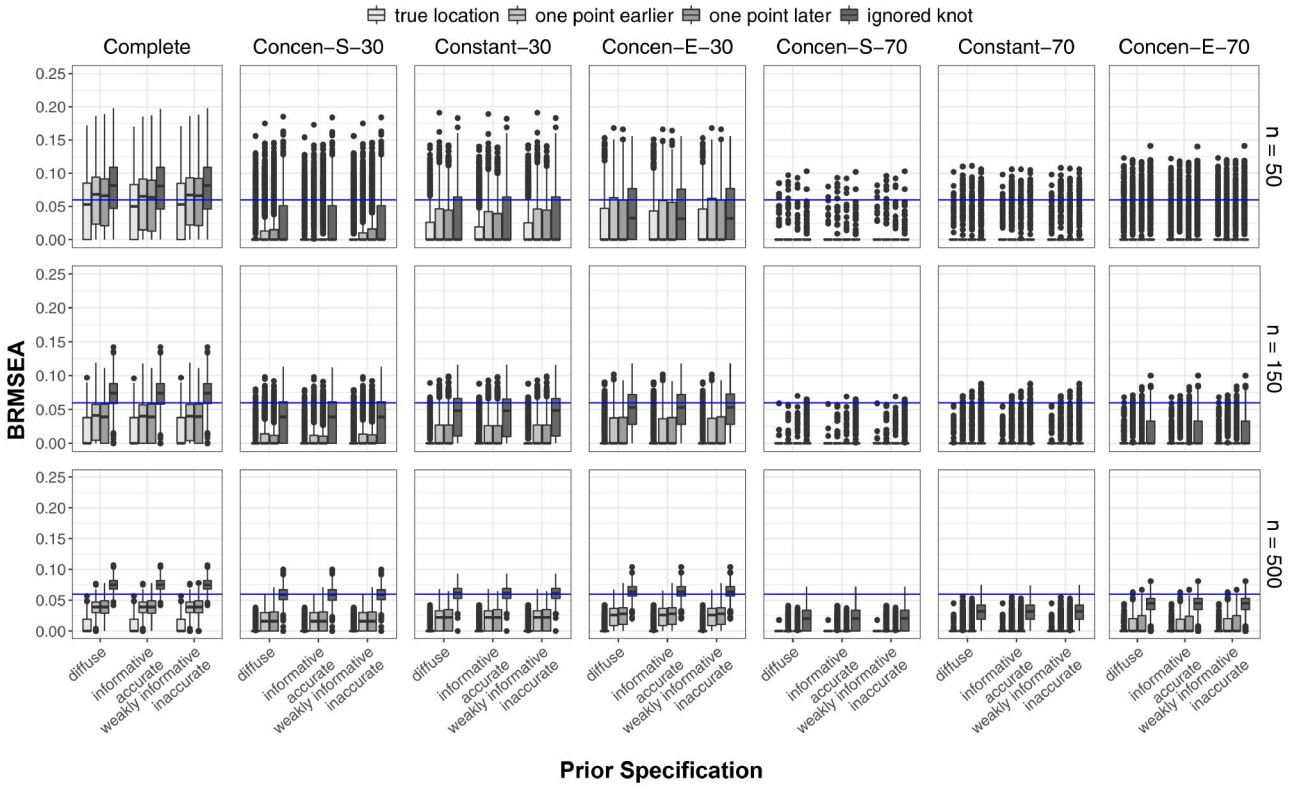

Figure 5. The BRMSEA across simulation conditions.

well-performing in detecting model misspecification when boxplots are located below 0.95 for misspecified models only.

Overall, the BCFI did not exhibit reliable performance in detecting misspecified models. When there were no missing data, the BCFI could only detect models with the ignored knot when $n = 500$. When the amount of missing data increased to 30%, the ability of the BCFI in detecting misspecified models with the ignored knot started to be worsened even when the sample size was 500. When 70% of data were missing, the BCFI became a completely unreliable measure to detect model misspecification regardless of sample sizes. These patterns were more severe when missingness was concentrated at the beginning. These findings indicate that the BCFI is vulnerable to missing data. In addition, the performance of the BCFI showed no much difference across prior specifications.

7.1.3. BRMSEA

We display results for the BRMSEA in Figure 5. All boxplots contain a solid horizontal line at .06 as a cutoff to determine whether a model fits data well or not: BRMSEA values $\leq$0.06 indicate a good model fit. If the BRMSEA performs well in detecting model misspecification, boxplots of misspecified models should be located above .06 whereas those of the correctly specified model below 0.06.

Our findings suggested that the BRMSEA was not a reliable measure to detect model misspecification. When data were complete, the BRMSEA obviously detected misspecified models with the ignored knot when the sample size was 150 or 500; still, models with misplaced knot locations could be detected. When the amount of missing data was 30%, the BRMSEA lost its ability to clearly detect misspecified models that ignored knot. Between three different attrition patterns, no difference was observed in the performance of the BRMSEA. When the percentage of missing data was 70%, the BRMSEA could not identify either correctly or misspecified models regardless of sample sizes, with this pattern being extreme when missing data arise in the early stages of the time points. As was the case with the BCFI (and BTLI), the little impact of different prior specifications was observed across all the simulation conditions.

7.2. Assessing Model Fit Using 90% CIs

Another method to assess model fit is to use 90% CIs for Bayesian approximate fit indices. Using CIs considers the whole posterior distributions instead of point estimates so that we can see plausible values of fit indices. We labeled three categories—good (conclusive), inconclusive, or poor (conclusive)—to classify fit and created stacked bar plots in Figures 6 and 7. Within each figure, there is a multitude of stacked bar plots. The rows represent sample sizes, the columns correspond to each level of the missing data condition, the x-axis represents model specifications, and the y-axis refers to the proportion of replications that belong to either conclusively good, inconclusive, or conclusively poor classification. Model fit is classified as conclusively good if the entire 90% CI is located within the" good" range, such that BCFI $\geq$0.95, BTLI $\geq$0.95, or BRMSEA $\leq$0.06. Model fit is defined conclusively poor if the corresponding 90% CI belong to BCFI $<$0.95, BTLI $<$0.95, or BRMSEA $>$0.06.

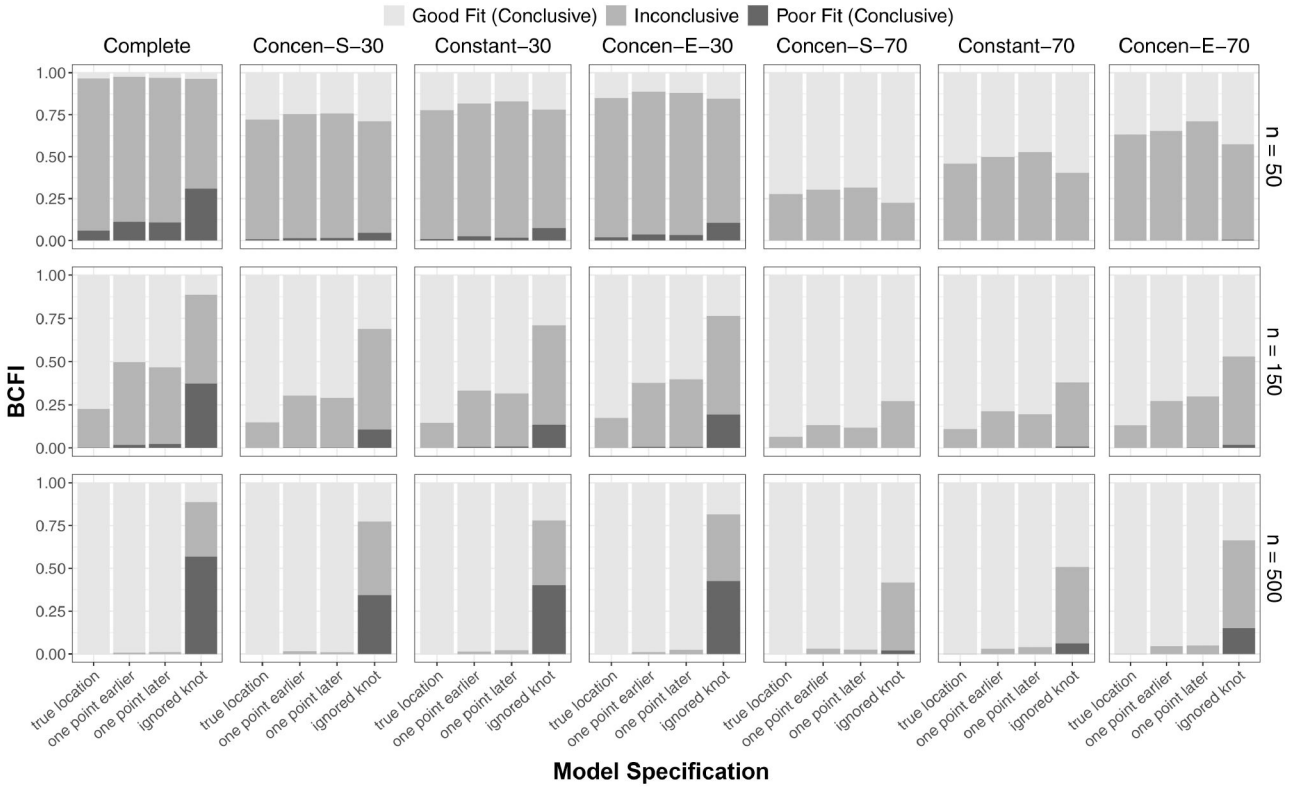

Figure 6. Classification based on the 90% credible interval of BCFI across simulation conditions.

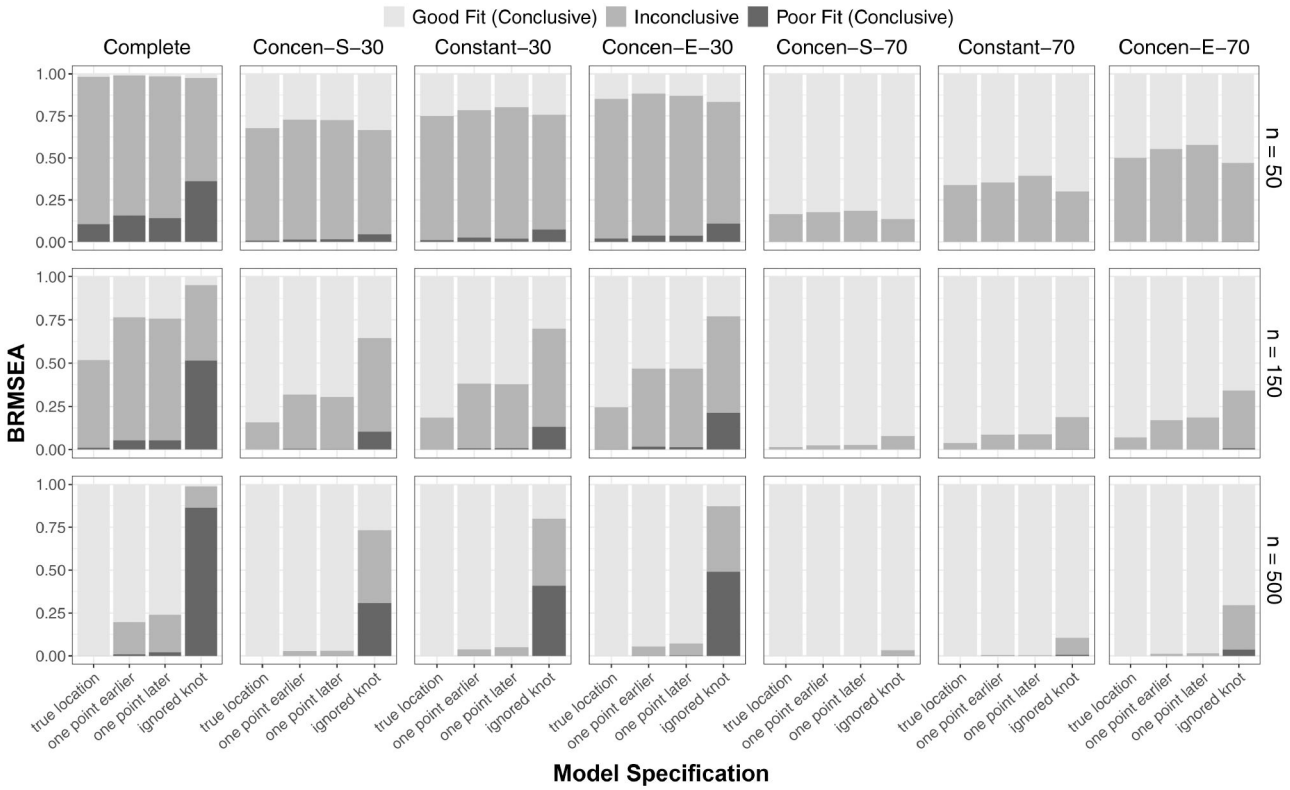

Figure 7. Classification based on the 90% credible interval of BRMSEA across simulation conditions.

Model fit is classified as inconclusive when the interval contains the cutoff value. In our initial analysis, results showed little difference across different prior specifications; thus, we collapsed results across prior specification conditions.

7.2.1. BCFI (and BTLI)

Figure 6 showcases classification based on the 90% CIs of the BCFI. When there were no missing data, most classifications were either inconclusive or conclusively poor with

$n = 50$. As sample size increased, we observed a higher proportion of replications classified as conclusively good. For $n = 500$, almost all replications for the true model were classified as conclusively good; however, misspecified models with misplaced knots were also classified as conclusively good. The ignored knot condition, on the other hand, still retained higher rates of replications classified as conclusively poor. This overall pattern was observed across the other levels of the missing data condition. For conditions with 70% of missing data, even a higher percentage of replications falsely classify misspecified models as conclusively good; such a pattern was worse when missing values were concentrated at the beginning. Based on these findings, using 90% CIs of BCFI did not work in reliably detecting model misfit.

7.2.2. BRMSEA

Results of the 90% CIs of the BRMSEA are displayed in Figure 7. The global pattern was similar to that of the BCFI (and BTLI). However, we focus on some aspects to highlight observed differences. When there were no missing data, a higher percentage of replications was classified as inconclusive for misspecified models with misplaced knots when $n = 500$, compared to the BCFI. When data consisted of 70% of missing values, more replications were wrongly classified as conclusively good for the misspecified model with the ignored knot. This indicates that the BRMSEA cannot detect severe model misspecification in the presence of a large amount of missing data. Overall, the use of 90% CIs of BRMSEA did not perform well to detect model misspecifications.

7.3. Assessing Model Selection

Successful model selection should lead to selecting the correctly specified model over misspecified models. We created Tables 1 and 2 to evaluate the performance of Bayesian model selection indices based on values of the BIC and DIC. Seven main horizontal blocks represent the levels of the missing data condition, and three main vertical blocks compare the true model to three misspecified models, respectively. Within each block, rows represent sample sizes, and columns correspond to three different prior specifications. The numbers in the table reflect both selection rates and differences in the information criteria. First, each number represents the percentage of replications where the true model was preferred over one of the misspecified models based on values of the information criteria. Second, we boldfaced each number if an average of differences in the information criteria between the true model and one of the misspecified models is greater than 5. This way we provide how many replications preferred the true model and whether such preferences are supported by a bigger magnitude of differences.

7.3.1. BIC

Table 1 displays selection rates for the BIC. Across all missing data conditions, there was a general pattern that selection rates were higher when the true model was compared to misspecified models with misplaced knots than the misspecified model with the ignored knot. We also detected the effect of sample sizes such that the selection rates increased as sample sizes became larger, accompanied by larger discrepancies in information criteria. When the amount of

Table 1. Selection rates of the BIC: a comparison between the true model and misspecified models.

	True model vs. one-point earlier			True model vs. one-point later			True model vs. ignored KNOT		
n	DIF	I-A	WI-INA	DIF	I-A	WI-INA	DIF	I-A	WI-INA
					Complete				
50	71.7	71.8	71.7	66.3	66.1	66.8	**12.7**	**12.8**	**12.6**
150	**87.2**	**87.2**	**87.2**	**85.5**	**85.5**	**85.5**	56.7	56.7	56.5
500	**97.5**	**97.5**	**97.5**	**97.4**	**97.4**	**97.4**	**100.0**	**100.0**	**100.0**
					Concen-S-30				
50	63.4	63.7	63.4	62.5	62.4	63.2	**5.4**	**5.4**	**5.3**
150	80.9	81.0	80.9	81.5	81.6	81.6	35.9	36.2	35.9
500	**97.5**	**97.5**	**97.5**	**95.2**	**95.2**	**95.2**	**98.4**	**98.4**	**98.4**
					Constant-30				
50	68.0	68.5	67.8	66.5	66.6	67.5	**7.0**	**7.1**	**7.0**
150	83.1	83.4	83.1	82.0	82.0	82.1	38.2	38.3	38.2
500	**96.6**	**96.6**	**96.6**	**96.9**	**96.9**	**96.8**	**99.2**	**99.2**	**99.2**
					Concen-E-30				
50	68.2	68.3	68.0	66.1	66.2	66.8	**7.3**	**7.5**	**7.4**
150	81.4	81.3	81.3	84.5	84.5	84.5	41.8	41.8	41.8
500	**97.4**	**97.4**	**97.4**	**97.6**	**97.6**	**97.6**	**99.6**	**99.6**	**99.6**
					Concen-S-70				
50	48.5	52.3	49.3	67.6	65.5	67.7	**0.6**	**0.5**	**0.6**
150	73.4	73.9	73.6	73.7	73.8	74.0	**6.8**	**6.9**	**6.8**
500	**93.2**	**93.2**	**93.2**	**88.5**	**88.7**	**88.5**	60.2	60.3	60.2
					Constant-70				
50	53.8	56.0	53.2	74.9	73.0	75.2	**1.3**	**1.4**	**1.3**
150	78.1	78.2	77.9	74.5	74.7	74.8	**9.2**	**9.3**	**9.2**
500	**93.1**	**93.1**	**93.2**	**92.5**	**92.4**	**92.5**	**76.0**	**76.3**	**76.1**
					Concen-E-70				
50	59.0	59.9	58.3	74.6	73.3	75.6	**1.8**	**1.6**	**1.8**
150	78.7	78.7	78.7	81.6	81.6	81.7	**15.5**	**15.5**	**15.4**
500	**94.4**	**94.4**	**94.3**	**96.2**	**96.3**	**96.2**	**90.6**	**90.6**	**90.6**

Note. DIF is the diffuse prior. I-A is the informative-accurate prior. WI-INA is the weakly informative-inaccurate prior. Numbers indicate the percentage of replications that favored the true model over the misspecified model. If the difference in information criteria between the true model and misspecified model was greater than 5, numbers were bold.

Table 2. Selection rates of the DIC: a comparison between the true model and misspecified models.

	True model vs. one-point earlier			True model vs. one-point later			True model vs. ignored knot		
n	DIF	I-A	WI-INA	DIF	I-A	WI-INA	DIF	I-A	WI-INA
					Complete				
50	70.1	68.7	69.4	64.7	64.4	65.2	62.3	65.1	63.2
150	86.5	86.5	86.5	85.1	85.1	85.1	**98.3**	**98.3**	**98.3**
500	**97.5**	**97.5**	**97.5**	**97.3**	**97.3**	**97.3**	**100.0**	**100.0**	**100.0**
					Concen-S-30				
50	65.6	64.9	65.0	60.5	59.1	60.8	52.4	56.1	52.9
150	80.6	80.2	80.5	81.0	80.9	81.2	**93.1**	**93.5**	**93.1**
500	**97.0**	**97.0**	**97.0**	**95.3**	**95.3**	**95.3**	**100.0**	**100.0**	**100.0**
					Constant-30				
50	69.0	69.0	68.8	63.7	62.8	63.9	56.6	60.3	57.3
150	83.4	83.3	83.4	82.1	81.5	82.2	**95.4**	**95.6**	**95.4**
500	**96.6**	**96.6**	**96.6**	**97.0**	**96.9**	**96.9**	**100.0**	**100.0**	**100.0**
					Concen-E-30				
50	70.2	69.5	69.7	63.5	62.0	63.7	57.3	62.1	57.9
150	81.9	81.8	81.9	83.2	82.9	83.2	**95.3**	**95.4**	**95.3**
500	**97.1**	**97.1**	**97.1**	**97.6**	**97.6**	**97.6**	**100.0**	**100.0**	**100.0**
					Concen-S-70				
50	59.3	60.5	58.3	59.8	55.7	60.7	36.4	42.4	36.4
150	75.1	75.2	75.1	69.9	69.2	69.9	72.5	74.8	72.2
500	**92.4**	**92.5**	**92.3**	**88.9**	**88.9**	**88.8**	**99.5**	**99.5**	**99.5**
					Constant-70				
50	61.1	62.2	60.2	65.2	63.1	65.5	41.6	46.9	41.7
150	79.1	79.1	78.8	71.4	70.1	71.6	**79.6**	**80.9**	**79.3**
500	**92.9**	**92.9**	**92.9**	**92.4**	**92.3**	**92.6**	**99.9**	**99.9**	**99.9**
					Concen-E-70				
50	65.9	66.6	65.0	65.2	62.3	65.2	45.2	50.8	45.6
150	80.1	80.1	79.9	77.4	76.8	77.2	**87.7**	**88.1**	**87.6**
500	**93.6**	**93.6**	**93.5**	**96.4**	**96.3**	**96.4**	**100.0**	**100.0**	**100.0**

Note. DIF is the diffuse prior. I-A is the informative-accurate prior. WI-INA is the weakly informative-inaccurate prior. Numbers indicate the percentage of replications that favored the true model over the misspecified model. If the difference in information criteria between the true model and misspecified model was greater than 5, numbers were bold.

missing data increased to 30% (i.e., the second through the fourth main horizontal block), selection rates decreased across all the conditions. When the amount of missing data was 70% (the fifth through the seventh main horizontal block), we observed an additional decrease in overall selection rates, indicating the presence of missing data negatively affected the performance of the BIC. Here, we noticed that when missing data were concentrated at the start of the study, selection rates were conspicuously lower compared to the other two conditions. We further observed the negligible effect of different prior specifications on the performance of the BIC across all the conditions.

7.3.2. DIC

Selection rates for the DIC are presented in Table 2. With complete data, the DIC could mostly detect models with misplaced knots for $n = 500$ and those with the ignored knot for $n = 150$ and $n = 500$. As the amount of missing data increased to 30% and further to 70%, we found decreases in selection rates across all conditions. However, when n equaled 500, the DIC still reliably favored the true model regardless of different types of model misspecification. When 70% of data were missing, we noticed that selection rates were considerably lower when missing data were concentrated at the beginning of the study compared to the other two attrition patterns, particularly when $n = 50$. This finding indicates the negative effect of a large amount of missing data arising at the start of the study when sample sizes are small. Regarding the impact of different prior specifications, there were no conspicuous differences between prior specification conditions.

7.3.3. Comparison Between BIC and DIC

When the knot was misplaced, the performances between these two selection indices in detecting model misfit were comparable. However, we found outperformance of the DIC over the BIC when the true model was compared to the misspecified model with the ignored knot. Particularly, when the sample sizes were 50 or 150, the true model selection rates of the DIC were substantially higher than those of the BIC. These patterns were observed across different missing data conditions. The DIC is thus preferred to the BIC in detecting a severe form of model misspecification.

8. A Secondary Simulation Study

The population model in the simulation setups described earlier was a linear-linear PGM with a single knot. Yet, it is important to acknowledge the possibility of utilizing more complex PGMs. One potential avenue is a PGM with multiple knots to capture transitions between more than just two distinct growth phases (e.g., Chung et al., 2017; Harring et al., 2021; Kroese et al., 2013). To explore the performance of Bayesian model evaluation tools in more complex piecewise growth modeling scenarios, we carried out a secondary simulation study using a linear-linear-linear PGM with two knots as a data-generating model. To maintain conciseness in the design of this secondary simulation study, we

considered a smaller number of simulation conditions and used only default prior specifications in M*plus* (Muthén & Muthén, 2017).

8.1. Simulation Design

This secondary simulation study was implemented with a fully crossed factorial design with three design factors: sample sizes (3 levels), missing data (5 levels), and knot placement (4 levels), resulting in a total of 60 cells. Each cell consisted of 1,000 replications.

8.1.1. The Population Model

The data-generating model was a linear-linear-linear PGM with seven repeated measurements. To formulate a PGM in which three linear growth segments join at two knots at occasions k_1 and k_2 with $k_1 < k_2$, we can express y_{ij} in Equation (2) as follows:

$$y_{ij} = \eta_{0i} + \eta_{1i} \times \min(t_j, k_1) + \eta_{2i} \times \min(\max(0, t_j - k_1), k_2 - k_1) + \eta_{3i} \times \max(t_j - k_2, 0) + \epsilon_{ij},$$

where η_{3i} refers to the third latent linear slop. For the population model, we placed knot locations at the third ($k_1 = 2$) and fifth ($k_2 = 4$) time points. Latent means of the three slopes were 0.5, 0.75, and 1 to indicate large changes from the first to the second slope and from the second to the third slope in terms of effect sizes (Cohen, 2002; Raudenbush & Liu, 2001). The path diagram with population parameter values is presented in Figure 8.

8.1.2. Sample Size

We included the following 3 levels of the sample size condition: 50 (small), 150 (medium), and 500 (large).

8.1.3. Missing Data

We considered the following 5 levels of the missing data condition: no missing values (Complete), missing values concentrated at the start of the study with 30% in the end (Concen-S-30), missing values concentrated at the end of the study with 30% in the end (Concen-E-30), missing values concentrated at the start of the study with 70% in the end (Concen-S-70), and missing values concentrated at the end of the study with 70% in the end (Concen-E-70).

8.1.4. Knot Placement

We included the following 4 levels of the knot placement condition: both knots were placed at the true locations

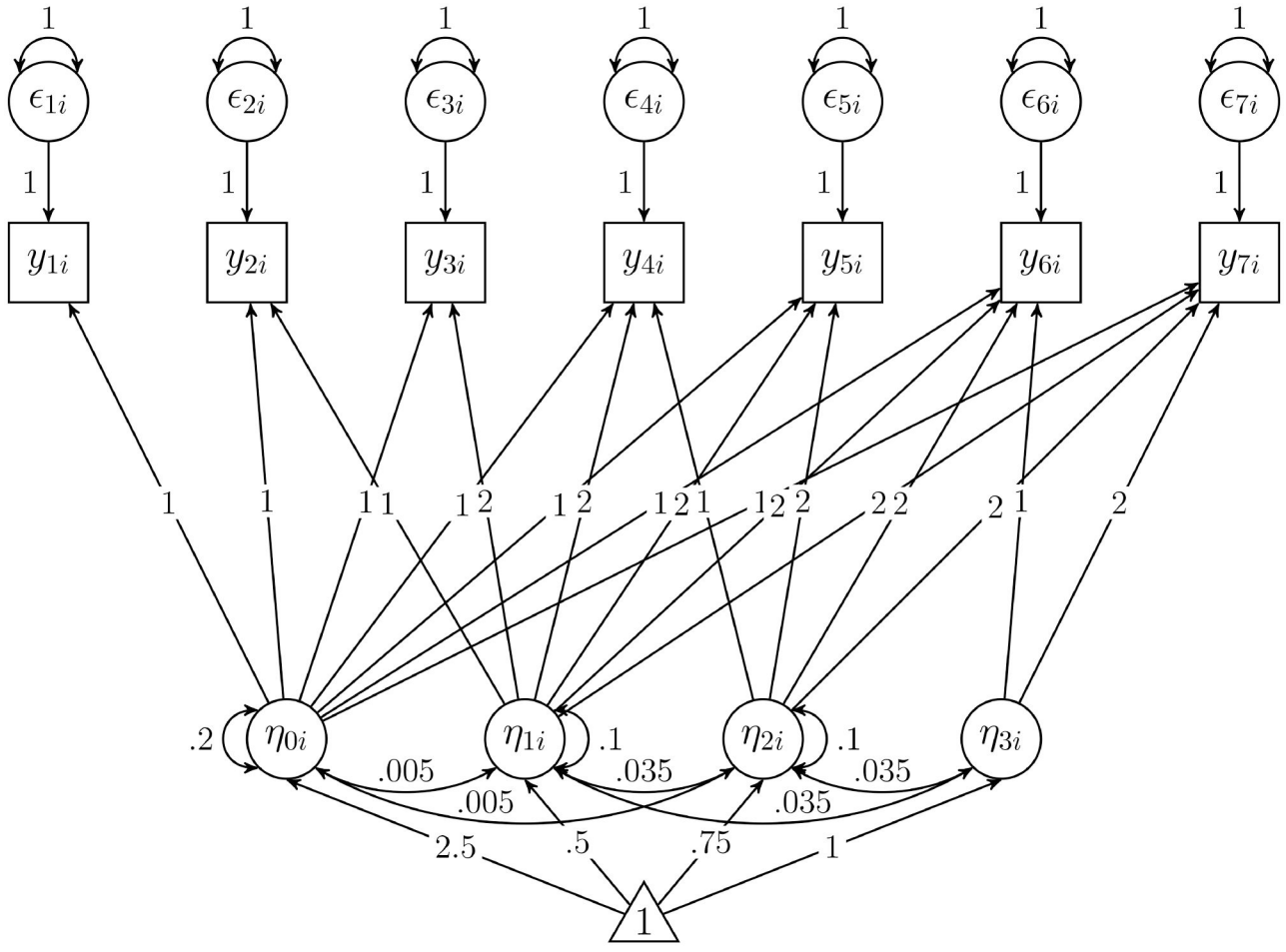

Figure 8. A population model of the secondary simulation study.

(True Location), one knot was correctly placed at the third time point and the other was ignored (One Correct Knot), one knot was misplaced at the fourth time point and the other was ignored (One Incorrect Knot), and two knots were entirely ignored (Ignored Knots). When one knot location was ignored (i.e., One Correct Knot and One Incorrect Knot), a linear-linear PGM was fitted. When both knot locations were ignored (i.e., Ignored Knots), a linear latent growth model was fitted.

8.1.5. Software and Bayesian Estimation

We used the R `lavaan` package (Rosseel, 2012) for data generation and used M*plus* version 8.6 (Muthén & Muthén, 2017) for Bayesian estimation. We implemented the Gibbs sampler to generate two MCMC chains each consisting of 20,000 iterations and had the thinning interval of 1. The first 10,000 iterations were discarded as burn-in samples. To assess chain convergence, we checked $\hat{R}$ statistics (Vehtari et al., 2021), with their values less than 1.05 as the criterion for convergence.

8.1.6. Outcomes of Interest

For Bayesian model fit indices (i.e., PPP-value, BRMSEA, BCFI, and BTLI), we extracted their values. For Bayesian model selection indices (i.e., BIC and DIC), we computed the proportion of replications that favored the true model over misspecified models. The difference in values of the BIC and DIC was also computed.

8.2. Results

On average, 92.6% of replications converged with $\hat{R}$ diagnostics for all parameters less than 1.05. The median percentage of converged replications was 100%. Because of similar results between the BCFI and BTLI, we mainly provide the BCFI results and attach the BTLI results as Online Supplementary Materials. The ways the figures and tables are interpreted are the same as explained in the previous Results section.

8.2.1. PPP-Value

We provide the results for the PPP-value in Figure 9 and begin by looking into conditions when there were no missing values. For $n = 50$, the PPP-value failed to detect the true model but could detect misfit for the model with one correct knot and the model with ignored knots. However, the PPP-value could not detect misfit for the model with one incorrect knot. When the sample size increased to 150, the PPP-value showed improved performance in detecting the true model but still could not detect misfit for the model with one incorrect knot. For the sample size of 500, the PPP-value performed well in detecting model misfit because the PPP-values of the true model hovered around .5 whereas those of the other three misspecified models hovered around or less than .05. As the proportion of missing data increased to 30%, we observed a similar pattern, except the PPP-value could not detect misfit for the model with ignored knots when the sample size was 50. When 70% of the data were missing, the performance of the PPP-value became worsened. For $n = 50$, the PPP-value could not

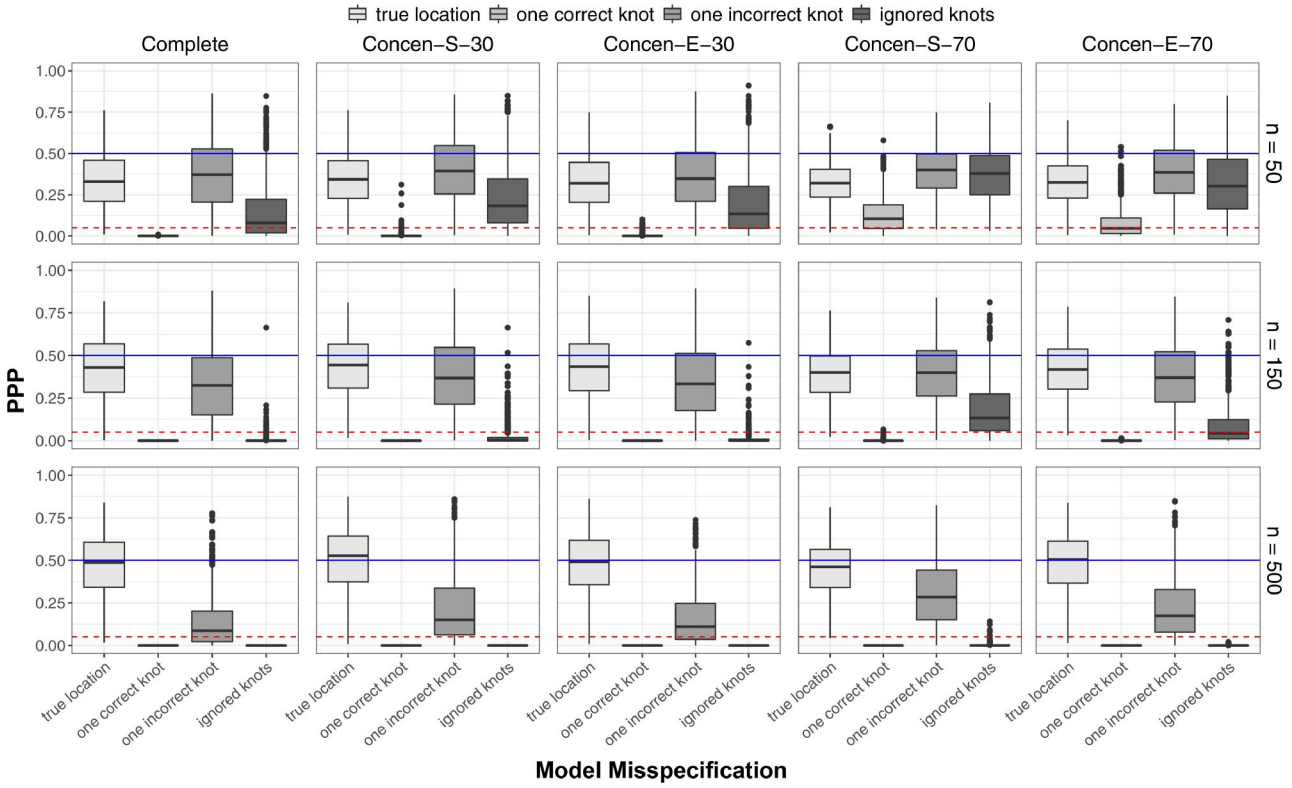

Figure 9. The PPP-value across simulation conditions of the secondary simulation study.

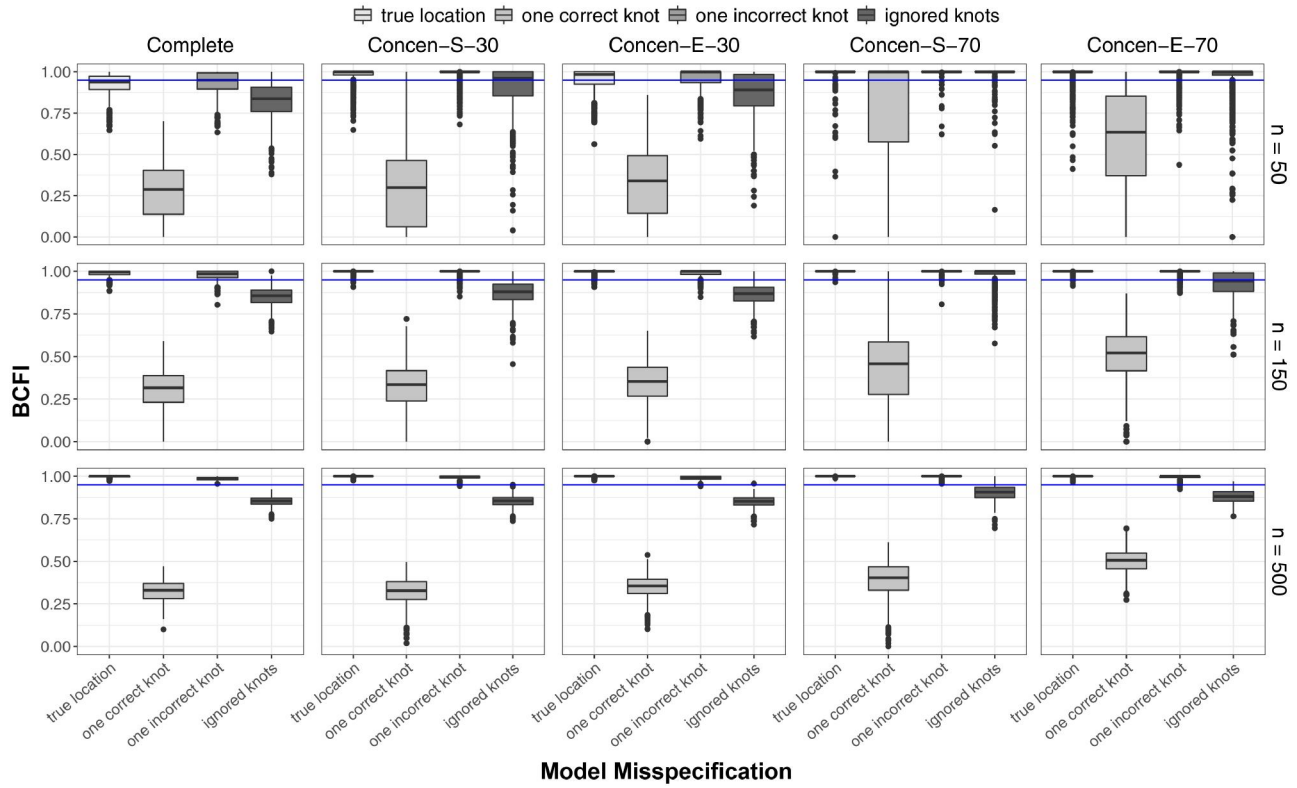

Figure 10. The BCFI across simulation conditions of the secondary simulation study.

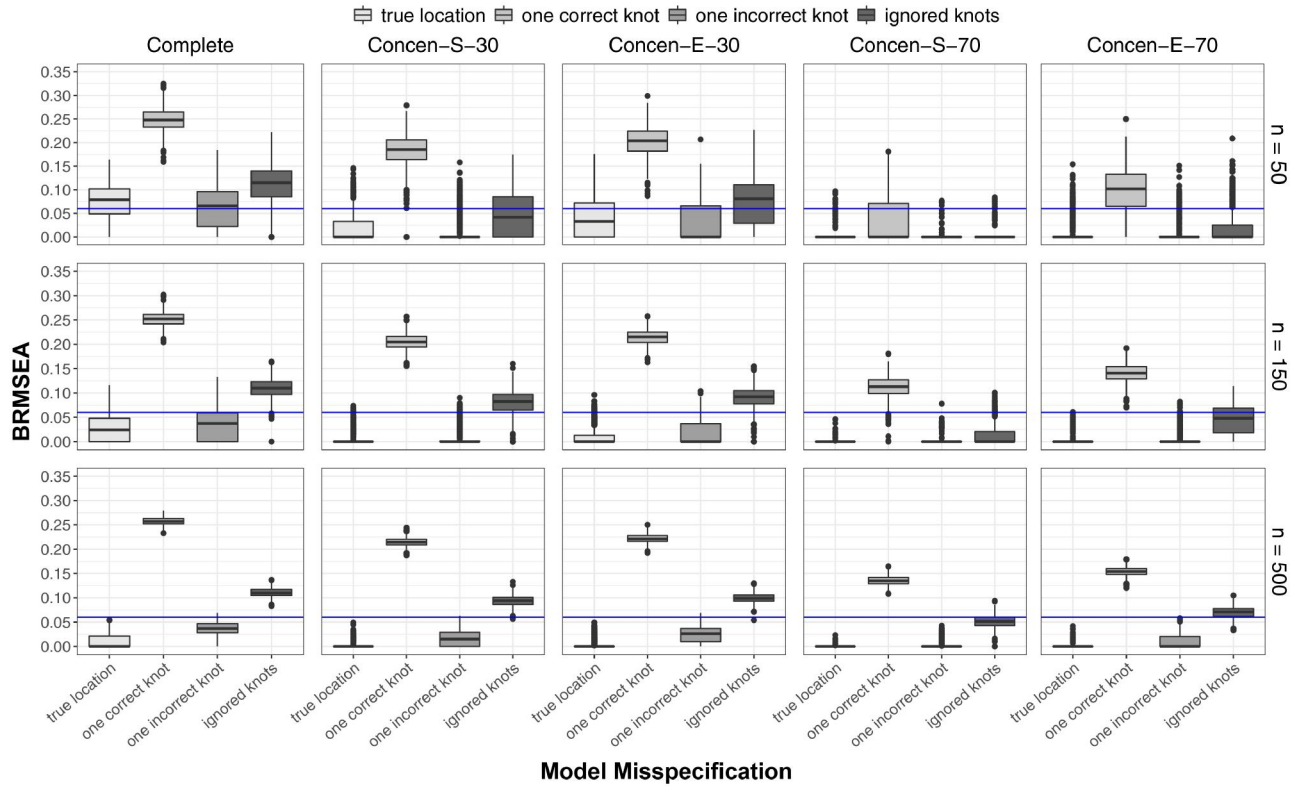

Figure 11. The BRMSEA across simulation conditions of the secondary simulation study.

detect the true model and failed to detect misspecified models with one incorrect knot or with ignored knots. When n increased to 150, the PPP-value started to detect misfit for the model with ignored knots. For $n = 500$, the PPP-value performed well in detecting the true model and identifying misfit for the models with one correct knot or with ignored knots. However, misfit for the model with one incorrect knot was not detected.

8.2.2. BCFI (and BTLI)

Figure 10 presents the results for the BCFI. In the absence of missing values and for $n = 50$, the BCFI did not reliably indicate a good fit for the true model. However, with larger sample sizes (150 or 500), the BCFI showed a good fit for the true model, but this was also the case for the model with one incorrect knot. When the amount of missing data was 30%, a similar pattern was observed. For $n = 50$, however, the BCFI could not detect misfit for the model with ignored knots. When the proportion of missing data was 70%, the overall performance was even worse for the sample sizes of 50 and 150, because the BCFI did not detect misfit for the model with one incorrect knot and for the model with ignored knots. For $n = 500$, misfit was still not detected for the model with one incorrect knot.

8.2.3. BRMSEA

The results for the BRMSEA are displayed in Figure 11. When there were no missing values, with the sample size of 50, the BRMSEA could not detect the true model. As the sample size increased to 150 and 500, the BRMSEA correctly indicated a good fit for the true model. For the misspecified models, the BRMSEA detected model misfit for the model with one correct knot and the model with ignored knots but indicated a good fit for the model with one incorrect knot. With 30% missing data, we found a similar pattern, except for the $n = 50$ condition. For the sample size of 50, the BRMSEA started to indicate a good fit not only for the true model but also for the models with one incorrect knot or with ignored knots. When the proportion of missing data was 70%, the BRMSEA performed worse. For $n = 50$ and the Concen-S-70 condition, the BRMSEA indicated a good fit to all four models, failing to detect any model misfit. When the sample size was 150 or 500, the BRMSEA could detect misfit for the model with one correct knot. In particular, for $n = 500$ and the Concen-E-70 condition, the BRMSEA additionally detected misfit for the model with ignored knots.

8.2.4. BIC and DIC

The selection rates for the BIC and DIC are provided in Table 3. We begin with focusing on the similarities between the BIC and DIC. Across all missing data conditions, the selection rates were the highest when the true model was compared with the model with one correct knot and the lowest when the true model was compared with the model with one incorrect knot. In addition, the selection rates decreased as the amount of missing data was larger. Another global pattern observed was that the selection rates tended to increase with larger sample sizes. However, this increase was relatively smaller when comparing the true model with the model with one incorrect knot. We additionally note the Concen-S-70 condition that showed the lowest selection rates for each comparison between the true model and misspecified models. This suggests that a larger amount of missing data occurring at the beginning of the study had a detrimental impact on the performance of the BIC and DIC in favoring the true model over misspecified models.

Turning to the differences between the BIC and DIC, we observed that the DIC generally had higher selection rates than the BIC. This means that the DIC performed better in detecting model misfit in all comparisons between the true model and misspecified models. In particular, when the true model was compared with the model with one incorrect knot, the DIC performed substantially better than the BIC when the sample sizes were 500. These findings indicate that the DIC is preferred to the BIC in detecting model misfit when complex PGMs are used.

9. Discussion

The Bayesian PGM is a useful modeling framework to describe longitudinal data exhibiting segmented developmental phases. Successful implementation of Bayesian PGMs depends on model evaluation, which is influenced by two methodological issues: model misspecification and missing data. A limited understanding of these issues prompted the current investigation. Our goal has been to understand the

Table 3. Selection rates of the BIC and DIC of the secondary simulation study: a comparison between the true model and misspecified models.

		True location vs. one correct knot			True location vs. one incorrect Knot			True location vs. ignored knots		
	n	Complete	Concen-S-30	Concen-S-70	Complete	Concen-S-30	Concen-S-70	Complete	Concen-S-30	Concen-S-70
	50	**100.0**	**99.2**	36.8	**0.0**	**0.0**	**0.0**	**2.7**	**0.9**	**0.0**
	150	**100.0**	**100.0**	**99.8**	**0.4**	**0.2**	**0.1**	56.8	**25.4**	**0.8**
BIC	500	**100.0**	**100.0**	**100.0**	**7.4**	**2.2**	**0.5**	**100.0**	**99.7**	59.0
	n		Concen-E-30	Concen-E-70		Concen-E-30	Concen-E-70		Concen-E-30	Concen-E-70
	50		**99.7**	65.3		**0.1**	**0.1**		**1.1**	**0.0**
	150		**100.0**	**100.0**		**0.3**	**0.0**		35.4	**3.7**
	500		**100.0**	**100.0**		**3.3**	**2.0**		**99.9**	**91.0**
	n	Complete	Concen-S-30	Concen-S-70	Complete	Concen-S-30	Concen-S-70	Complete	Concen-S-30	Concen-S-70
	50	**100.0**	**100.0**	**98.8**	20.1	20.1	17.9	**72.2**	61.5	33.7
	150	**100.0**	**100.0**	**100.0**	47.6	38.6	28.7	**99.8**	**99.6**	**84.5**
DIC	500	**100.0**	**100.0**	**100.0**	**90.0**	**83.8**	64.9	**100.0**	**100.0**	**100.0**
	n		Concen-E-30	Concen-E-70		Concen-E-30	Concen-E-70		Concen-E-30	Concen-E-70
	50		**100.0**	**99.6**		24.0	18.9		63.3	40.6
	150		**100.0**	**100.0**		43.1	40.1		**99.4**	**94.6**
	500		**100.0**	**100.0**		**86.8**	**82.3**		**100.0**	**100.0**

Note. Numbers indicate the percentage of replications that favored the true model over the misspecified model. If the difference in information criteria between the true model and misspecified model was greater than 5, numbers were bold.

impact of these two factors on the performance of Bayesian model fit and selection indices, with an additional focus on their sensitivity under different prior specifications. In this section, we wrap up and discuss important findings, provide practical guidelines for researchers, and outline our future research directions.

9.1. Performance of Bayesian Model Fit and Selection Indices

Our first aim was to investigate how well the Bayesian model fit and selection indices could evaluate Bayesian PGMs in the presence of model misspecification and missing data. We assessed the performance of those indices under one-knot and two-knot PGM scenarios. One global pattern we found under the one-knot scenario was that those indices were better at detecting severely misspecified models with the ignored knot than mildly misspecified models with misplaced knots. This pattern is in accordance with Depaoli et al. (2023) in which the severe degree of model misspecification was more likely to be detected in Bayesian PGMs. The only exception to this pattern was the BIC, which showed poor performance in detecting severe misspecification when sample sizes were small or medium. When there were two knots, the issue became more complicated. In particular, when one knot was misplaced and simultaneously the other was ignored, misfit was hardly detected by model fit and selection indices. Such complication highlights the importance of recognizing the limitations of those indices utilized in complex PGM scenarios where multiple types of misfit can coexist. In addition, the abilities of indices in detecting model misfit were worsened when the amount of missing data increased. This is not surprising, in light of previous research that has shown issues with the accuracy of model fit and selection indices when dealing with a larger proportion of missing data (Asparouhov & Muthén, 2021; Winter & Depaoli, 2022b). What our finding extends is the role of attrition pattern. If a large amount of data were missing (i.e., 70% missingness at the last time point), there was an effect of attrition pattern such that the indices performed worse in detecting misfit when missing data were concentrated at the beginning of the study. This signals applied researchers that they are encouraged to carefully devise data collection plans to prevent higher dropouts in the initial stages of longitudinal studies.

Across different simulation conditions, sample size played an important role in the performance of various indices. With complete data, the PPP-value was a reliable index when sample sizes became larger; however, they appeared to be unreliable for smaller sample sizes. The increasing amount of missing data led to a loss of the ability of the PPP-value in detecting model misspecification. In the 70% missing data conditions, in particular, the PPP-value could not detect models with mild misspecifications for large sample sizes. When no missing data were present, the three Bayesian approximate fit indices (i.e., BCFI, BTLI, and BRMSEA) were less useful than the PPP-value in that, as sample size increased, they indicated a good fit for misspecified models. Caution is thus needed in using these Bayesian approximate fit indices, as what has been pointed out in methodological literature (Asparouhov & Muthén, 2021; Depaoli et al., 2023; Winter & Depaoli, 2022b). In addition, a larger amount of missing data could increase the degree to which these three approximate fit indices falsely favored the misspecified models. The approach to evaluating model fit based on the 90% CIs of the BCFI, BTLI, and BRMSEA showed limited utility; they could only detect severe model misfit when there were no missing data under the largest sample size conditions. As Winter and Depaoli (2022b); Depaoli et al. (2023) discussed, the unreliable performance of CI-based evaluation method can be attributable to other factors than model misspecification per se, and our study provides evidence that sample size is one factor that came into play.

Between the two Bayesian model selection indices (i.e., BIC and DIC), the DIC appeared to be reliable in detecting severely misspecified models for at least medium sample sizes compared to the BIC. This is alarming to the use of the BIC: Researchers are likely to reach erroneous conclusions if they ignore the true knots in growth models. Furthermore, for the two selection indices, the sample size of 50 is too small to conclusively detect wrong knot locations. The medium sample size at the very least is required to conclusively detect model misfit when no missingness is present.

9.2. Prior Sensitivity

The second aim regarded whether the performances of various indices were sensitive to different prior specifications. According to our simulation results, there was little impact of different prior settings. These results repeat what has been found in Depaoli et al. (2023) such that prior specifications were not a major component in detecting misspecified Bayesian PGMs. The use of the informative-accurate prior was not distinctively advantageous than the diffuse or weakly informative-inaccurate priors, although it is known that different prior specifications influence the ability of Bayesian model fit and selection indices to detect model misfit (e.g., Cain & Zhang, 2019; Winter & Depaoli, 2022b). One possible explanation for the negligible impact of prior distributions lies in how the population models are specified. The means of the first and second slope factors were 0.5 and 0.75 in our simulation study. Suppose, for instance, that the magnitude of the second growth phase becomes extremely larger (e.g., 7.5) for a true model, and a misspecified model with the ignored knot is fitted. Specifying informative-accurate priors for the mean of the second slope can aid in avoiding falsely favoring the misspecified model because the posterior mean estimate of the second slope will hardly overlap with the wrong model.

We therefore argue that the findings from the current simulation study should not be interpreted as a complete absence of the impact of prior distributions. Prior distributions can be specified in infinite ways according to researchers. The level of alignment with true values and

Table 4. Practical guidelines on using model fit and selection indices in Bayesian PGMs.

Index	Key recommendations
PPP-value	Use it to detect mild or severe misspecifications when sample sizes are large without any missing data. Avoid using it when sample sizes are small. Avoid using it when there is a higher proportion of missing data concentrated at the beginning of the study.
BCFI BTLI BRMSEA	Use them to detect severe misspecification when sample sizes are large without any missing data. Avoid using them when there is a higher proportion of missing data concentrated at the beginning of the study. Consider using them as model comparison tools.
BIC DIC	Use both to detect mild or severe misspecifications when sample sizes are large. Use both cautiously when there is a higher proportion of missing data concentrated at the beginning of the study. Use the DIC instead of the BIC when your analysis model does not have a knot and sample sizes are at least medium. Avoid using both when sample sizes are small.

informativeness of prior distributions can result in different conclusions when using the Bayesian model fit and selection indices (Winter & Depaoli, 2022b). Bayesian PGMs have other parameters that can be susceptible to different prior settings such as knot locations or covariance structures and can be extended to complex forms (e.g., Kohli et al., 2015; Lock et al., 2018). This insight implies that the choice of prior settings for other parameters in different forms of Bayesian PGMs can have a greater influence on model evaluation, potentially altering the performance of model fit and selection indices.

9.3. Practical Guidelines for Researchers

The ideal scenario in piecewise growth modeling is without model misspecification and missing data, as they are often the sources of misleading conclusions. However, in reality, both issues are prevalent. We therefore provide Table 4 as a comprehensive set of guidelines for the utilization of Bayesian model assessment tools to aid applied researchers in their use.

In addition, given the less desirable performance of the Bayesian approximate fit indices (i.e., BCFI, BTLI, and BRMSEA) in assessing a single model, we recommend that researchers consider using these indices as a toolbox for model comparison when multiple models are compared (see also Depaoli et al., 2023; Winter & Depaoli, 2022a, 2022b for similar arguments). This approach is particularly useful given their limited utility in reliably detecting model misspecifications in the presence of missing data. To demonstrate, suppose a researcher is uncertain about which model is optimal among multiple candidate models with knot locations at different time points, and is interested in using the BRMSEA to evaluate the models. In this case, the researcher can use the BRMSEA as a model comparison index to identify the most optimal model that has the smallest value among multiple candidate models. This way, the researcher can still incorporate information from the BRMSEA without relying solely on model fit assessment.

9.4. Future Research Directions

In the current simulation design, we focused on the misspecification in the mean structure, particularly trajectories and knot placements. While these types of misspecifications drew methodological interests (Depaoli et al., 2023; Kwok et al., 2010; Leite & Stapleton, 2011; Ning & Luo, 2017), we point out the misspecified covariance structure as another possible form of misspecification. Covariance structures can be misspecified by, for example, fixing the variances of or covariances between growth factors to zero (see also, e.g., Wu et al., 2009; Wu & West, 2010; Winter & Depaoli, 2022b). The application of PGMs can carry misspecification in the covariance structure, and this could also have a substantial impact on the performance of Bayesian model fit and selection indices, which is a direction that future research can dive into.

A research topic that has not received enough attention is the performance of the Bayes factor as a model selection index. The Bayes factor is the ratio of the marginal likelihoods of two models under comparison and quantifies which model is better supported by observed data over the other (Kass & Raftery, 1995). We did not include the Bayes factor in the current study because it is not provided as default in *Mplus*. One of the desirable characteristics of the Bayes factor is its ability to penalize complex models that have large parameter space (Jefferys & Berger, 1992). We thus find its potential advantage in detecting misfit of complex models such as Bayesian PGMs in the presence of missing data.

Another potential topic regards missing data mechanisms. Our study generated MAR data, which is reasonable for analyzing longitudinal data (Enders, 2011). It is possible, however, that data could have been generated from a missing data mechanism other than the MAR assumption. When missingness is nonignorable (i.e., MNAR), the performance of Bayesian model assessment measures is expected to be different. For instance, Shin et al. (2017) found that the performance of the PPP-value was worsened under the MNAR assumption than the MCAR or MAR assumptions. Future works can direct the focus to the effect of the MNAR assumption on Bayesian PGMs and provide guidelines regarding how researchers can prevent undesirable model selection outcomes.

9.5. Concluding Remarks

Evaluating Bayesian PGMs is crucial to reach correct and valid research conclusions, and model misspecification and missing data are two important factors that can affect model evaluation. We examined the performance of model fit and selection indices in Bayesian PGMs and provided guidelines

on how to wisely use these indices in the presence of model misspecification and missing data. We are hopeful that our guidelines can aid future researchers interested in Bayesian piecewise growth modeling.

Disclosure statement

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

ORCID

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

References

Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Wiley.

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

Asparouhov, T., & Muthén, B. O. (2010). Bayesian analysis using Mplus: Technical implementation. Technical Report.

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

Blozis, S. A. (2007). On fitting nonlinear latent curve models to multiple variables measured longitudinally. *Structural Equation Modeling*, *14*, 179–201. https://doi.org/10.1080/10705510709336743

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

Can, S., van de Schoot, R., & Hox, J. (2015). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement*, *75*, 406–427. https://doi.org/10.1177/0013164414547959

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*, 651–673. https://doi.org/10.1214/06-BA122

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. (2002). *Statistical power analysis for the behavioral sciences*. Routledge.

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 and G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 3–24). IAP Information Age Publishing.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and hall/CRC.

Depaoli, S., Jia, F., & Heo, I. (2023). Detecting model misspecifications in Bayesian piecewise growth models. *Structural Equation Modeling*, *30*, 574–591. https://doi.org/10.1080/10705511.2022.2144865

Diallo, T. M. O., Morin, A. J. S., & D, P. P. (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

Edwards, K. D., & Konold, T. R. (2023). Impact of informative priors on model fit indices in Bayesian confirmatory factor analysis. *Structural Equation Modeling*, *30*, 272–283. https://doi.org/10.1080/10705511.2022.2126359

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16. https://doi.org/10.1037/a0022640

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*, 76–93. https://doi.org/10.1037/met0000102

Fay, D. M., Levy, R., & Schulte, A. C. (2022). Model criticism of growth curve models via posterior predictive model checking. *The Journal of Experimental Education*, *90*, 191–212. https://doi.org/10.1080/00220973.2020.1711697

Flora, D. B. (2008). Specifying piecewise latent trajectory models for longitudinal data. *Structural Equation Modeling*, *15*, 513–533. https://doi.org/10.1080/10705510802154349

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*, 46–70. https://doi.org/10.1037/met0000224

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd edition). CRC Press Boca.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016. https://doi.org/10.1007/s11222-013-9416-2

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

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development*, *82*, 1357–1371. https://doi.org/10.1111/j.1467-8624.2011.01630.x

Gustavson, K., von Soest, T., Karevold, E., & Røysamb, E. (2012). Attrition and generalizability in longitudinal studies: Findings from a 15-year population-based study and a Monte Carlo simulation study. *BMC Public Health*, *12*, 918. https://doi.org/10.1186/1471-2458-12-918

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. https://doi.org/10.3758/s13428-020-01420-5

Hoofs, H., van de Schoot, R., Jansen, N. W., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement*, *78*, 537–568. https://doi.org/10.1177/0013164417709314

Hu, B. Y., Fan, X., Wu, Y., LoCasale-Crouch, J., & Song, Z. (2020). Teacher–child interaction quality and Chinese children's academic and cognitive development: New perspectives from piecewise growth modelling. *Early Childhood Research Quarterly*, *51*, 242–255. https://doi.org/10.1016/j.ecresq.2019.10.003

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

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, *48*, 370–397. https://doi.org/10.1080/00273171.2013.778191

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

Kohli, N., Peralta, Y., & Bose, M. (2019). Piecewise random-effects modeling software programs. *Structural Equation Modeling*, *26*, 156–164. https://doi.org/10.1080/10705511.2018.1516507

Kroese, F. M., Adriaanse, M. A., Vinkers, C. D., van de Schoot, R., & de Ridder, D. T. (2013). The effectiveness of a proactive coping intervention targeting self-management in diabetes patients. *Psychology & Health*, *29*, 110–125. https://doi.org/10.1080/08870446.2013.841911

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

Lee, S.-Y. (2007). *Structural equation modelling: A Bayesian approach*. John Wiley & Sons.

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, F., Duncan, T. E., & Hops, H. (2001). Examining developmental trajectories in adolescent alcohol use using piecewise growth mixture modelling analysis. *Journal of Studies on Alcohol*, *62*, 199–210. https://doi.org/10.15288/jsa.2001.62.199

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. https://doi.org/10.3389/fpsyg.2019.00583

Lock, E. F., Kohli, N., & Bose, M. (2018). Detecting multiple random changepoints in Bayesian piecewise growth mixture models. *Psychometrika*, *83*, 733–750. https://doi.org/10.1007/s11336-017-9594-5

Lu, Z. L., & Zhang, Z. (2021). Bayesian approach to non-ignorable missingness in latent growth models. *Journal of Behavioral Data Science*, *1*, 1–30. https://doi.org/10.35566/jbds/v1n2/p1

Lu, Z. L., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research*, *46*, 567–597. https://doi.org/10.1080/00273171.2011.589261

Marcoulides, K. M. (2018). Automated latent growth curve model fitting: A segmentation and knot selection approach. *Structural Equation Modeling*, *25*, 687–699. https://doi.org/10.1080/10705511.2018.1424548

Muthén, B. O. (2010). Bayesian analysis in Mplus: A brief introduction. Technical Report.

Muthén, L. K., & Muthén, B. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide*. Muthén & Muthén.

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, *41*, 143–153. https://doi.org/10.1177/01650254156 18275

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

Ortega-Azurduy, S., Tan, F., & Berger, M. (2008). The effect of dropout on the efficiency of d-optimal designs of linear mixed models. *Statistics in Medicine*, *27*, 2601–2617. https://doi.org/10.1002/sim.3108

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. https://doi.org/10.1037/a0021939

Ram, N., & Grimm, K. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development*, *31*, 303–316. https://doi.org/10.1177/0165025407077751

Raudenbush, S. W., & Liu, X.-F. (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. https://doi.org/10.1037/1082-989X.6.4.387

Rhemtulla, M., Jia, F., Wu, W., & Little, T. D. (2014). Planned missing designs to optimize the efficiency of latent growth parameter estimates. *International Journal of Behavioral Development*, *38*, 423–434. https://doi.org/10.1177/0165025413514324

Rosseel, Y. (2012). lavaan: An R package for structural equation modelling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

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

Serang, S., Zhang, Z., Helm, J., Steele, J. S., & Grimm, K. J. (2015). Evaluation of a Bayesian approach to estimating nonlinear mixed-effects mixture models. *Structural Equation Modeling*, *22*, 202–215. https://doi.org/10.1080/10705511.2014.937322

Shi, D., DiStefano, C., Zheng, X., Liu, R., & Jiang, Z. (2021). Fitting latent growth models with small sample sizes and non-normal missing data. *International Journal of Behavioral Development*, *45*, 179–192. https://doi.org/10.1177/0165025420979365

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*, 310–334. https://doi.org/10.1177/0013164418783530

Shin, T., Davison, M. L., & Long, J. D. (2017). Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychological Methods*, *22*, 426–449. https://doi.org/10.1037/met0000094

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*, 583–639. https://doi.org/10.1111/1467-9868.00353

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540. https://doi.org/10.1080/01621459.1987.10478458

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology*, *55*, 329–337. https://doi.org/10.1016/s0895-4356(01)00476-0

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*, 667–718. https://doi.org/10.1214/20-BA1221

Winter, S. D., & Depaoli, S. (2022a). Performance of model fit and selection indices for Bayesian structural equation modelling with missing data. *Structural Equation Modeling*, (*0*), 1–22. https://doi.org/10.1080/10705511.2022.2164286

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

Wu, W., & Jia, F. (2021). Applying planned missingness designs to longitudinal panel studies in developmental science: An overview. *New Directions for Child and Adolescent Development*, *2021*, 35–63. https://doi.org/10.1002/cad.20391

Wu, W., Jia, F., Rhemtulla, M., & Little, T. D. (2016). Search for efficient complete and planned missing data designs for analysis of change. *Behavior Research Methods*, *48*, 1047–1061. https://doi.org/10.3758/s13428-015-0629-5

Wu, W., & West, S. G. (2010). Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behavioral Research*, *45*, 420–452. https://doi.org/10.1080/00273171.2010.483378

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. https://doi.org/10.1037/a0015858

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes [Doctoral dissertation, University of California, Los Angeles]. ProQuest Dissertations & Theses Global.

Zhang, Z., Hamagami, F., Lijuan 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/10.1177/0165025407 077764