UNIVERSITY OF CALIFORNIA, MERCED

Advanced Methods for Detecting Specification Issues in Bayesian Structural Equation Modeling

A dissertation for the degree
Doctor of Philosophy

in

Psychological Sciences

by

Sonja D. Winter

2021

Committee members:
Professor Sarah Depaoli, Chair
Professor Fan Jia
Professor Keke Lai
Professor Jack Vevea

Signature Page

The dissertation of Sonja D. Winter is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

(Professor Fan Jia)

_____

(Professor Keke Lai)

_____

(Professor Jack Vevea)

_____

(Professor Sarah Depaoli, Chair)

University of California, Merced
2021

To my parents

Table of Contents

Online Supplemental Material:

Supporting R code and supplemental figures can be found on the Dissertation OSF page: https://osf.io/xyaqk/.

List of Figures

List of Tables

Acknowledgements

I have been supported by countless people during my five years at UC Merced. I feel so lucky to have become part of such an encouraging academic family that was there for me throughout the ups and downs of this thing called graduate school.

First, I would like to thank my academic advisor, Dr. Sarah Depaoli. From the day that we talked about Bayes over soup in Amsterdam I knew you were a powerhouse in our field and someone I would love to by mentored by. Thank you for accepting me into your lab. I would not have achieved my academic successes without your guidance. Beyond teaching me about the nuts and bolts of quantitative methods and research, you have also taught me so much about succeeding as a woman in academia. Thank you for being open with me and showing me how to overcome the challenges that appeared on my path. I hope we will keep working together for many years to come!

I would also like to thank the members of my dissertation committee, Dr. Jack Vevea, Dr. Keke Lai, and Dr. Fan Jia. You were each so generous in sharing your expertise and helping me understand the finer points of things such as mathematical formulas, grammar, communicating through visualizations, and designing high quality, rigorous simulation studies. And, although not on any official committee, many other members of the Psychological Sciences faculty have also given me guidance throughout my time at UC Merced. I would like to thank Dr. Jan Wallander, Dr. Anna Song, Dr. Heather Bortfeld, Dr. Deborah Wiebe, Dr. Haiyan Liu, Dr. Ren Liu, Dr. Matthew Zawadzki, and Dr. Jitkse Tiemensma for teaching me important lessons about research, academia, and professional and personal growth.

I would not have succeeded in my graduate studies without the friendship and incredible support of my lab mates. I have learned so much from those who came before me (June, Patrice, and Johnny) and am learning so much more from those who are in the lab now (Marieke and Lydia). More importantly, I could (and can) always talk to anyone about anything that was bothering me and get sympathy and wisdom in return. I also want to shout out returning visiting lab member Sanne Smid, who graciously shared some essential code for my dissertation with me. Likewise, I want to mention my adoptive lab mates: Katie and Susette, and most importantly, Amber. Thank you for always sharing your lab with windows! And Amber, thank you so, so much for talking to me on the bus to Yosemite during recruitment weekend back in 2016 and being the best work wife I could have ever wished for. Amber, together with Larisa, Tammy, and Kaylyn were also my Merced roomies who provided a safe haven to unwind and commiserate over risotto and wine. I am so excited to see what amazing things you will all accomplish!

Someone who has been by my side (quite literally this past year) from day one is my husband, Funs. Not only did he volunteer to be my personal tech support, writing coach, and practice audience, but he also gave me the space to focus on my dissertation without any distractions (hmmm, maybe besides Nina). I would not have been able to finish this dissertation without him. And finally, I want to thank my parents, to whom I dedicate this dissertation. Even though we live on different continents and time zones, you were always there to love me and support me unconditionally, even when you had no idea what it was that I was doing! I hope I have made you proud.

VITA

EDUCATION

2016 – Present **Ph.D. in Psychological Sciences**
Dissertation: Advanced Methods for Detecting Specification Issues in Bayesian Structural Equation Modeling
Graduate Advisor: Dr. Sarah Depaoli

2011 – 2013 **Master of Science in Developmental Psychology**
University of Utrecht, Utrecht, The Netherlands
Thesis: What matters? The role of cognitive competence and affective beliefs in educational achievement during adolescence.
Thesis Advisor: Dr. Jan Boom

2008 – 2011 **Bachelor of Science in Developmental Psychology**
University of Utrecht, Utrecht, The Netherlands
Thesis: Differential susceptibility? Interaction of the DRD4-7R polymorphism and parenting practices on aggressive and prosocial behaviour.
Thesis Advisor: Dr. Geertjan Overbeek

SELECT PUBLICATIONS

Depaoli, S., **Winter, S. D.**, & Visser, M. (2020). The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. *Frontiers in Psychology*, *11*(November), 1–18. https://doi.org/10.3389/fpsyg.2020.608045

Smid, S. C., & **Winter, S. D.** (2020). Dangers of the Defaults: A Tutorial on the Impact of Default Priors When Using Bayesian SEM With Small Samples. *Frontiers in Psychology*, *11*(December), 287–290. https://doi.org/10.3389/fpsyg.2020.611963

Winter, S. D., & Depaoli, S. (2019). An illustration of Bayesian approximate measurement invariance with longitudinal data and a small sample size. *International Journal of Behavioral Development*, *44*(4), 371-382. https://doi.org/10.1177/0165025419880610

Abstract of the Dissertation

Advanced Methods for Detecting Specification Issues in Bayesian Structural Equation Modeling

by
Sonja D. Winter

Doctor of Philosophy in Quantitative Methods, Measurement, and Statistics

University of California, Merced, 2021
Professor Sarah Depaoli, Chair

This dissertation consists of two studies investigating model and prior specification issues in the context of Bayesian structural equation modeling (SEM). Two of the major advantages of Bayesian estimation for SEM are that complex models can more easily be estimated, and prior information can be directly included in the analysis. Two aspects of Bayesian estimation of SEM that are important for the applied researcher are model and prior specification assessment. In Study 1 of this dissertation, I examined the ability of several model fit and selection indices to detect model misspecification in two commonly used SEMs with data that are completely observed or that contain missing values. Simulation results showed that Bayesian approximate model fit indices may not be appropriate for model fit assessment of a single model. The posterior predictive *p*-value was more likely to detect model misspecification, although it was sensitive to sample size and the presence of missing values. Instead of focusing on a single model, researchers should aim to compare multiple models and focus on model selection, using Bayesian approximate fit indices, in addition to model fit assessment. Furthermore, informative priors that diverge from the population model worsened model fit even for a correctly specified model. Thus, researchers should examine whether there is disagreement between the priors and their observed data when using informative priors. This so-called prior-data disagreement was the focus of Study 2 of this dissertation. In this study, I examined three indices for detecting prior-data disagreement, the Data Agreement Criterion (DAC), Bayes Factor (BF), and prior-predictive *p*-value, and assessed their ability to detect diverging priors across 4 sample sizes and 49 prior specifications for the mean intercept and linear slope parameters of a latent growth model. Simulation results showed that while the DAC was easily implemented, it cannot assess interactions between priors placed on different parameters. Use of the BF becomes unfeasible as model complexity, sample size, or the number of prior specifications examined increase. Here, the prior-predictive *p*-value may offer an alternative, although it may not be appropriate for prior specifications that are partially or fully diffuse. Furthermore, all prior-data disagreement indices tended to be better at detecting disagreement with larger sample sizes, whereas the impact of disagreement is largest with small sample sizes. Other implications, suggestions for applied researchers, limitations, and future directions are also discussed.

Chapter 1

Overview of Dissertation

Bayesian estimation of structural equation modeling (SEM) is becoming an increasingly popular approach in the social sciences. This increasing popularity is not surprising given that Bayesian estimation allows a researcher to estimate complex models more easily and incorporate prior information into the analysis. Accurate model and prior specification may be the most important parts of Bayesian estimation to explore further because they are crucial to the integrity of the use of Bayesian SEM. In this dissertation, I investigate both factors to provide researchers with concrete recommendations for model and prior specification assessment. I will discuss various Bayesian indices of model fit and selection and assess their ability to detect model misspecification across a variety of SEMs. Similarly, I will discuss three indices for detecting disagreement between the prior specification and the observed data and assess their ability to detect prior-data disagreement in a commonly used SEM. I am interested in these indices' capability to support applied researchers in making informed decisions about Bayesian SEMs.

This dissertation consists of two separate studies. The first study is entitled "Performance of Model Fit and Selection Indices for Bayesian Structural Equation Modeling". The primary goal of Study 1 is to examine the ability of the Bayesian information criterion (BIC), deviance information criterion (DIC), posterior-predictive *p*-value (PPP), and Bayesian root mean square error of approximation (BRMSEA), comparative fit index (BCFI), and Tucker-Lewis index (BTLI) to detect model misspecification. This study will focus on three different SEMs and evaluate the impact of (a) sample size, (b) missing at random (MAR) data in one or multiple variables, (c) location and severity of the misspecification, and (d) prior specification.

The second study is entitled "Detecting Prior-Data Disagreement in Bayesian Structural Equation Modeling". In Study 2, the main goal is to investigate the ability of the data agreement criterion (DAC), Bayes factor (BF), and prior-predictive *p*-value to detect prior-data disagreement in an SEM. This study focuses on the latent growth model, an SEM for which researchers are likely to specify informative priors on the main intercept and growth parameters. The indices will be compared across 49 different prior specifications that represent increasing amounts of prior-data disagreement. In addition to assessing whether the indices can detect prior-data disagreement, I will also evaluate whether the prior-data disagreement causes bias in the posterior estimates.

This dissertation is structured as follows. First, I will provide a general introduction to the Bayesian statistics that are the focus of the two studies. Next, I will present two studies aiming to: (1) examine a variety of Bayesian model fit and selection indices and assess their ability to detect model misspecification across a variety of SEMs (Study 1 of the dissertation), and (2) examine three indices for detecting prior-data disagreement in a commonly used SEM (Study 2 of the dissertation). Finally, I will conclude by discussing the implications of my findings regarding model- and prior-specification and provide recommendations for use in applied research settings.

Chapter 2

General Introduction to Bayesian Statistics

The use of Bayesian statistical methods has been increasing in psychological science (van de Schoot et al., 2017) since their introduction to the field in the 1960s (Edwards et al., 1963; Rupp et al., 2004). This chapter will introduce the Bayesian estimation framework and discuss its application to structural equation model (SEM) estimation.

2.1 Bayesian Estimation

2.1.1 Bayes' Theorem

Bayes' theorem was first introduced in the 18th century (Bayes, 1764; Laplace, 1774) to use conditional probability to express how the probability of an event is updated by the availability of prior evidence. Bayes' theorem is stated mathematically as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{1}$$

where A and B are events, and $P(B) \neq 0$. Further, $P(A|B)$ is the likelihood of event A occurring given that B is true and $P(B|A)$ is the likelihood of event B occurring given that A is true. Finally, $P(A)$ and $P(B)$ are the marginal probabilities of event A and B occurring, respectively. Using the Bayesian interpretation, these probabilities express a degree of belief before and after event B is observed. $P(A)$ is the prior, or initial, degree of belief in A. $P(A|B)$ is the posterior degree of belief in A, after incorporating the knowledge that event B is true. The support provided by event B for event A is represented in the quotient $\frac{P(B|A)}{P(B)}$. Typically, event B is treated as fixed as it refers to the observed data. In this case, the posterior probability is proportional to the numerator of Bayes' theorem, or the prior multiplied by the likelihood:

$$P(A|B) \propto P(A)P(B|A). \tag{2}$$

2.1.2 Bayesian Estimation

Bayes' theorem can be used for Bayesian estimation, a statistical estimation method in which a probability distribution of subjectively likely values (prior) is updated with new evidence (data likelihood). In general, statistical estimation is used to obtain estimates for unknown parameters, $\boldsymbol{\theta}$, given the data $\mathbf{y}$. How the unknown parameters are defined is something that differentiates Bayesian estimation from frequentist estimation. In

frequentist estimation, the assumption is that θ is unknown but fixed. In Bayesian estimation, θ is assumed to be random, with a probability distribution that quantifies the uncertainty about the true value of θ. Thus, in the context of Bayesian estimation, we can express Bayes' theorem in terms of a vector of parameters, $\boldsymbol{\theta}$ and a sample vector of data, $\mathbf{y}$, as follows:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3}$$

In Equation (3), $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, $p(\mathbf{y}|\boldsymbol{\theta})$ is the data likelihood for $\mathbf{y}$, or $L(\boldsymbol{\theta}|\mathbf{y})$, and $p(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution.

2.1.3 The Prior

Specifying prior distributions for all parameters $\boldsymbol{\theta}$ is the first step of a Bayesian analysis. Different ways of categorizing types of prior distributions have been proposed, such as subjective versus objective or informative versus non-informative. In both cases, the former implies a type of prior that is deliberately specified to convey prior knowledge, whereas the latter implies a type of prior that is deliberately specified to convey as little information as possible. In practice, the amount of information conveyed by any particular prior depends on many factors, such as the overall model, the parameter in question, and the sample size (Gelman et al., 2017; Smid & Winter, 2020; van Erp et al., 2018). Under certain circumstances, a prior that was assumed to be objective or non-informative may strongly affect the posterior distribution (e.g., Depaoli & Clifton, 2015; van Erp et al., 2018). However, the general purpose of non-informative priors is to "let the data speak for themselves" (Gelman et al., 2013). Thus, posteriors estimated with non-informative priors will typically reflect the information provided by the observed data.

To illustrate the interplay between the prior distribution and the data likelihood, I will discuss a simple example. Suppose we have a sample of twenty individuals for which we have observed a change in the number of depression symptoms from summer to winter (measured on a scale from -30 to 30). We are interested in two parameters: the mean and variance of the exam score. Those parameters will tell us to what extent the number of depression symptoms changed on average and to what extent the change in depression symptoms varied across individuals. For the current illustration, I will focus on the mean parameter.[1] Remember that, within the Bayesian framework, parameters are assumed to unknown and random, with a probability distribution that quantifies the uncertainty about the true value of the parameter. Typically, the likelihood of a mean parameter is set to follow a normal distribution. This likelihood is often combined with a prior that also follows a normal distribution (this choice is more fully discussed in Section 2.1.4).

[1] As the variance cannot be negative, its prior usually follows a distribution that is only defined for positive values, such as the inverse gamma distribution. Other distributions can be specified, such as the uniform, half-*t*, or half-Cauchy distributions. For the sake of simplicity, I do not include a discussion of this prior here and instead focus on the prior for the mean parameter.

Each prior distribution has its own parameters, which are called hyperparameters. For example, the normal distribution has a mean μ and variance σ^2. Taking an uninformative, or objective, approach to Bayesian estimation, we might specify a prior for the mean parameter that follows a normal distribution centered around 0 with a variance of 1000, which we can express as follows:

$$\mu_{\text{score}} \sim N(\mu = 0, \sigma^2 = 1000). \quad (4)$$

This prior distribution is depicted in Figure 1, panel A for a range of average change in depression symptoms from -10 to 30. As can be seen in panel A, this prior places almost equal prior probability on the entire range shown on the *x*-axis (i.e., the line is almost flat). The reason for the flat appearance of this normal distribution is the variance hyperparameter, which corresponds to a standard deviation of 100. As the prior is centered around 0, 68% of the prior distribution contains values between -100 and 100, and 95% of the prior distribution contains values between -200 and 200. With this prior, a wide range of values can be sampled from the posterior distribution of the average change in depression. Based on this prior, an average increase in depression symptoms of 15 is equally plausible as an average decrease in depression symptoms of 15.

Comparing the prior distribution to the data likelihood (shown in Figure 1, panel B) shows that the data likelihood provides more precise information about the value of the change in depression symptoms. The data-likelihood distribution more closely resembles a typical normal distribution, with most of its density placed around an average change in symptoms close to 15 and symmetric tails that indicate that average changes in symptoms of 0 or 30 are not plausible. As the data likelihood contains more information about the location of the average change in depression symptoms, the posterior distribution will closely resemble the data likelihood (panel C). This scenario is an example in which the prior "let the data speak for themselves" (Gelman et al., 2013).

In contrast, we could have specified an informative prior for the mean parameter. We could use previous groups of individuals' average change in depression symptoms from summer to winter as the basis for our prior belief. For example, we could look at the average change in depression symptoms of samples observed in the previous ten years, and find that the average change was 10, with a variance of 4. This information translates to the following prior:

$$\mu_{\text{score}} \sim N(\mu = 10,\ \sigma^2 = 4). \quad (5)$$

This informative prior distribution is depicted in Figure 1, panel D. Comparing panel D to panel A, we can see that the informative prior places more plausibility on values close to 10. For this prior, 95% of the distribution contains values between 6 and 14. Although any normal distribution has an infinite range, values close to 10 are much more likely to be sampled from the posterior distribution of the average change in depression symptoms with this informative prior. If we compare this prior to the data likelihood (shown again in Figure 1, panel E), we can see that the data likelihood distribution is only somewhat narrower and more peaked (i.e., it provides more precise information about the value of the average change in symptoms) than the prior distribution. Also, note that the

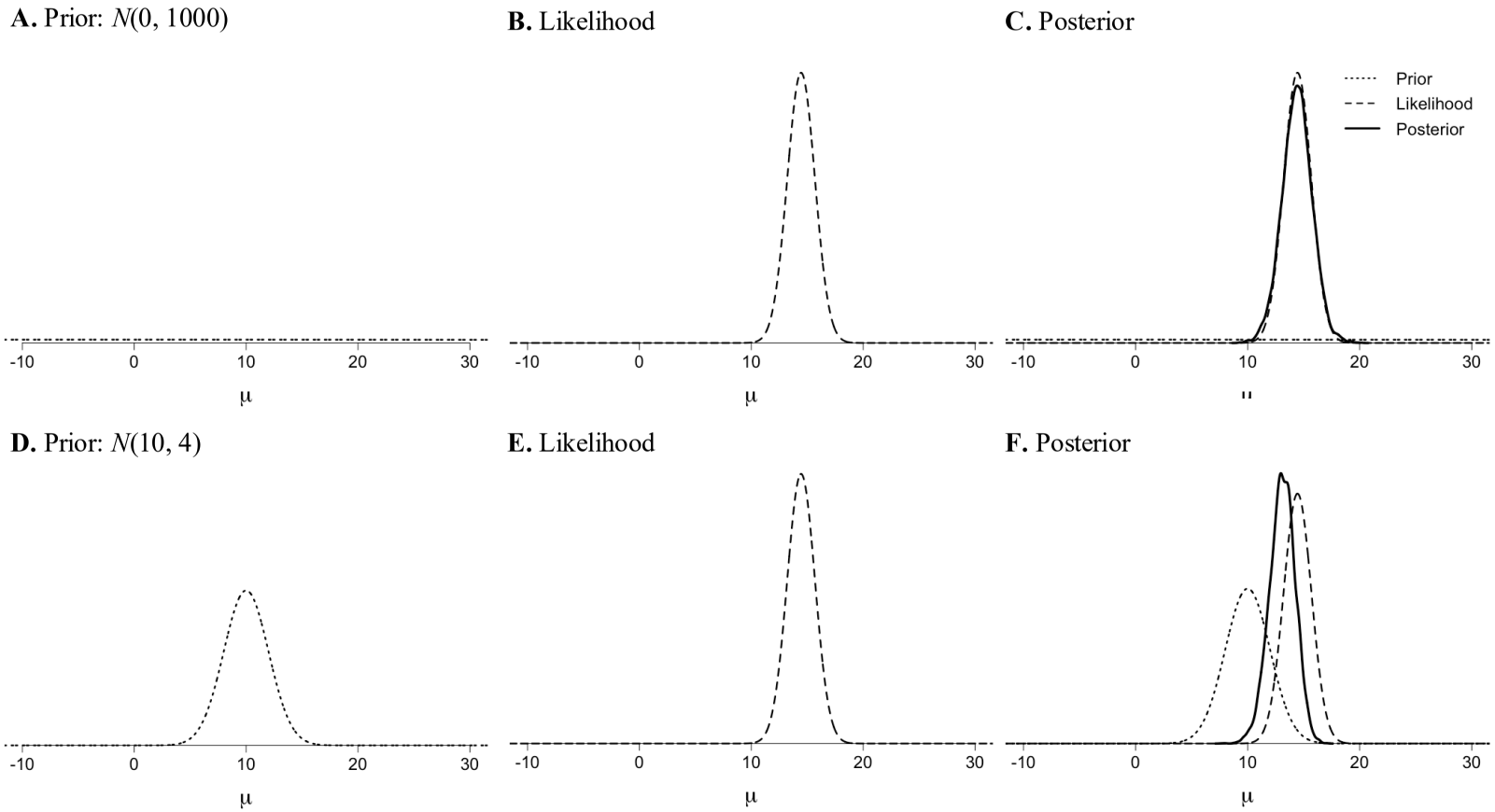


Figure 1. Illustrative example describing the interplay between the prior distribution, the likelihood, and the posterior distribution.

data likelihood remains the same regardless of the prior selected (i.e., panel B and panel E are identical). These middle panels represent the fixed information provided by the specific sample of twenty individuals through the data likelihood.

With an informative prior like the one specified in panel D, the posterior distribution will represent a compromise between these two sources of information. Indeed, the posterior distribution shown in Figure 1, panel F, falls between the prior and likelihood distributions, although it is somewhat closer to the likelihood distribution. This tendency towards the likelihood distribution makes sense, as the information provided by the likelihood is more precise (i.e., the density distribution is narrow) than the information provided by the prior. The difference between the posterior distributions in panel C and F demonstrate that the way a prior distribution is specified can affect the conclusions drawn from the posterior. With a diffuse prior, the posterior mean of the average change in depression symptoms is 14.43, but with the informative prior, the posterior mean drops to 13.07. Thus, for any analysis, it is important to consider the potential impact of the prior specification on the posterior estimates.

2.1.4 Computation

Prior to the 1990s, the application of Bayesian estimation was limited to problems and models that were relatively simple. The formulas presented above provide a closed-form expression only if the model includes a single parameter *and* if the posterior distribution of that parameter is in the same probability distribution family as the prior distribution. When both distributions come from the same probability distribution family, they are called conjugate distributions, and the prior becomes a conjugate prior for the likelihood function (Raiffa & Schlaifer, 1961; Wetherill, 1961). A well-known example is the Gaussian family, which is conjugate to itself. This means that if the likelihood function is Gaussian, choosing a prior distribution that is Gaussian will ensure that the posterior distribution is also Gaussian (as in Figure 1 above). Another advantage of conjugacy is that it makes it clear to see how the likelihood and the prior interact to generate the posterior.

Non-conjugate priors are prior distributions that are part of a different probability distribution family than the posterior distribution of a parameter. For example, the conjugate prior distribution for a variance parameter is the inverse gamma distribution. However, other distributions, such as the uniform, half-*t*, or half-Cauchy, can also be used as priors for this type of parameter (Gelman, 2006; Polson & Scott, 2012; van Erp et al., 2018). Researchers sometimes use non-conjugate priors because their distribution aligns better with the prior knowledge about the parameter of interest. In that case, numerical integration becomes necessary to obtain the posterior distribution. Numerical integration is also necessary when there are multiple parameters of interest. However, numerical integration can become untenable once the model includes more than a few parameters. This major computational challenge hampered the adoption of Bayesian estimation by the broader scientific community.

This changed with the introduction of the Markov chain Monte Carlo (MCMC; Hastings, 1970) estimation algorithm to Bayesian statistics (Gelfand & Smith, 1990; Smith & Roberts, 1993). MCMC methods generate systematic random samples from high-dimensional (e.g., across multiple parameters) probability distributions. Whereas

Monte Carlo sampling methods draw independent samples from a distribution, MCMC methods draw samples where the next sample is dependent on the existing sample, creating what is called a Markov chain. The advantage of these dependent samples is that the algorithm can efficiently narrow in on the parameter that is being approximated from the distribution. MCMC methods are not inherently Bayesian, but their application to approximating distributions aligns with the Bayesian definition of θ as a random variable with a probability distribution. To approximate a target distribution, MCMC methods evaluate integrals or sums through simulation instead of exact or approximate algebraic analysis. The general steps of the MCMC method are:

1. Obtain starting values $\boldsymbol{\theta^0}$.
2. Sample $\boldsymbol{\theta^1}$ using a specific algorithm or sampler.
3. Repeat step 2 *S* times to obtain a Markov chain $\{\boldsymbol{\theta^0}, \boldsymbol{\theta^1}, \dots, \boldsymbol{\theta^s}\}$.

The starting values used in Step 1 are placeholders that are simply used to start the MCMC process. As the initial samples in the Markov chain are still dependent on the starting values, they may not be a good representation of the (posterior) distribution that is being approximated. These initial samples are part of the *burn-in* or *warm-up* period, which is the period in the Markov chain before it enters a stationary distribution. Once a chain enters a stationary distribution, additional samples *s* will not alter the distribution (O Roberts, 1996). However, increasing the number of samples in the Markov chain will increase the precision with which the (posterior) distribution is estimated. Thus, samples that are part of the burn-in period are typically not included in the final set of samples to ensure that the portion of the Markov chain that remains is likely to represent the (posterior) distribution. While the explanation above focuses on estimating a single Markov chain, multiple Markov chains can be obtained for each parameter. By estimating multiple chains, it is possible to ensure that any single chain is not stuck in a particular sampling space (see Section 2.2.5 for an example).

The Markov chain (or chains) generated in Step 3 of the MCMC algorithm represents an estimate of the posterior distribution of θ. The obtained posterior distribution is usually summarized in various ways (see Section 2.2.6). As mentioned above, a Markov chain contains dependent samples. The dependence between samples within a Markov chain is influenced by a transition kernel. A transition kernel describes how parameter values are updated at each iteration *s* of the Markov chain (van de Schoot et al., 2021). The specific sampling algorithm used in Step 2 of the MCMC method determines the definition of the transition kernel. In the next two sections, I will introduce two sampling algorithms that can be used to generate the MCMC chains.

2.1.4.1 Gibbs Sampler

The Gibbs sampler (Geman & Geman, 1984) is probably the most well-known MCMC sampling algorithm. It is used as the default sampler in various Bayesian software programs and packages, such as the BUGS programming language, through WinBUGS (no longer supported; Gilks et al., 1994; Lunn et al., 2012) and OpenBUGS (Lunn et al., 2009), M*plus* (L. K. Muthén & Muthén, 2017), and (r)JAGS (Plummer, 2017, 2019). The Gibbs sampler first takes the set of starting values for all *q* model parameters, placed in a vector $\boldsymbol{\theta^0} = \left(\theta_1^0, \dots, \theta_q^0\right)'$. Next, the Gibbs sampler generates a new set of values for the

model parameters $\boldsymbol{\theta^s}$ from $\boldsymbol{\theta^{s-1}}$ for the Monte Carlo iterations $s = 1, 2, \ldots, S$ using the following steps:

$$
\begin{aligned}
&1.\ \text{sample}\quad \theta_1^s \sim p\left(\theta_1 \middle| \theta_2^{s-1}, \theta_3^{s-1}, \ldots, \theta_q^{s-1}, \mathbf{y}\right)\\
&2.\ \text{sample}\quad \theta_2^s \sim p\left(\theta_2 \middle| \theta_1^{s}, \theta_3^{s-1}, \ldots, \theta_q^{s-1}, \mathbf{y}\right)\\
&\vdots\\
&q.\ \text{sample}\quad \theta_q^s \sim p\left(\theta_q \middle| \theta_1^{s}, \theta_2^{s}, \ldots, \theta_{q-1}^{s}, \mathbf{y}\right).
\end{aligned}
$$

Thus, in the Gibbs sampler transition kernel, a new set of parameter values is generated based on their posterior conditional distribution, and the probability of accepting generated values is equal to one. After these steps are repeated for all *S* iterations, the resulting samples approximate the joint distribution of all parameters. Further, the marginal distribution of any subset of parameters can be approximated by considering the samples for that subset of variables, ignoring the rest. Finally, the expected value of any parameter (i.e., the posterior mean) can be approximated by averaging over all the (post *burn-in*) samples.

The samples resulting from the Gibbs sampler are Markov chains, which adhere to the Markov property: the future is independent of the past given the present state. A Markov chain is a discrete time stochastic process with the property that the distribution of $\boldsymbol{\theta^s}$ given all previous values of the process, $\boldsymbol{\theta^0}, \boldsymbol{\theta^1}, \ldots, \boldsymbol{\theta^{s-1}}$, depends only on $\boldsymbol{\theta^{s-1}}$ (O Roberts, 1996). Based on this property, we expect adjacent members from a Markov chain to be positively correlated, a phenomenon called autocorrelation. Autocorrelation can also exist between more distant members of the Markov chain. An important result regarding autocorrelation is that if the posterior samples are from a stationary process, correlated draws still provide an unbiased estimate of the distribution, provided that the sample size *S* is sufficiently large.

2.1.4.2 Other Samplers

While the Gibbs sampler may be the most well-known sampler, others exist. A recent addition to this lineup is the Hamiltonian Monte Carlo algorithm (HMC; Betancourt, 2018; Betancourt & Girolami, 2015; Neal, 2011) and its extension, the No-U-Turn sampler (NUTS; Hoffman & Gelman, 2014). These samplers rely on Hamiltonian dynamics in physics to efficiently sample from the posterior distribution. In essence, this algorithm treats the model as a high-dimensional particle that is moving across the posterior sampling space. According to the description given by the Stan Development Team (2020), HMC relies on auxiliary momentum variables ρ and draws samples from the following joint density:

$$p(\rho, \theta) = p(\rho|\theta)p(\theta). \tag{6}$$

In most implementations of HMC (e.g., Stan; Stan Development Team, 2020), the auxiliary density follows a multivariate normal distribution that does not depend on the parameters θ, $\rho \sim \mathrm{MVN}(0, \Sigma)$. The covariance matrix Σ works as a Euclidean metric to

rotate and scale the target distribution (for details, see: Betancourt & Stein, 2011). Further, Hamiltonian dynamics uses the joint density $p(\rho, \theta)$ to define a Hamiltonian:

$$\begin{aligned} H(\rho, \theta) &= -\log p(\rho, \theta) \\ &= -\log p(\rho|\theta) - \log p(\theta) \\ &= T(\rho|\theta) + V(\theta), \end{aligned} \tag{7}$$

where $\mathrm{T}(\rho|\theta) = -\log p(\rho|\theta)$ is the *kinetic energy* and $\mathrm{V}(\theta) = -\log p(\theta)$ is the *potential energy*. These terms are used to generate transitions from one space to another (i.e., move from sample $s - 1$ to sample s) in three steps. First, a value for the momentum ρ is drawn independently of the current parameter values (this means that momentum does not persist across iterations). Next, the joint system (θ, ρ), which is made up of the current parameter values and the new momentum value, is evolved using Hamilton's equations:

$$\begin{aligned} \frac{d\theta}{dt} &= +\frac{\partial H}{\partial \rho} = +\frac{\partial T}{\partial \rho} \\ \frac{d\rho}{dt} &= -\frac{\partial H}{\partial \theta} = -\frac{\partial T}{\partial \theta} - \frac{\partial V}{\partial \theta} = -\frac{\partial V}{\partial \theta}. \end{aligned} \tag{8}$$

To solve this two-state differential equation, most implementations of HMC use the leapfrog integrator. This numerical integration algorithm is specifically adapted to lead to stable results for Hamiltonian systems of equations. The leapfrog integrator takes discrete steps of some small interval ϵ (i.e., the step size). The leapfrog integrator alternates between half-step updates of the momentum and full-step updates of the parameter:

$$\begin{aligned} \rho &\leftarrow \rho - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta} \\ \theta &\leftarrow \theta + \epsilon \Sigma \rho \\ \rho &\leftarrow \rho - \frac{\epsilon}{2}\frac{\partial V}{\partial \theta}. \end{aligned} \tag{9}$$

The three leapfrogs steps above are repeated L times, resulting in a state that is denoted as (ρ^*, θ^*) (for details, see: Leimkuhler & Reich, 2004). Finally, a Metropolis acceptance step is applied, because numerical errors can occur during leapfrog steps. The probability of accepting the proposed (ρ^*, θ^*) is $\min\left(1, \exp\left(H(\rho, \theta) - H(\rho^*, \theta^*)\right)\right)$. If the proposed state is not accepted, the algorithm returns to the previous parameter values, which are used to initialize the next iteration. Thus, contrary to the Gibbs sampler, the NUTS transition kernel does not always accept the newly generated parameter value.

With HMC, each iteration in the chain can move a more considerable distance compared to the Gibbs sampler. This characteristic results in posterior samples that efficiently explore the entire posterior sampling space while promoting low dependence between iterations in the chain. Compared to the Gibbs sampler, the HMC algorithm and

the NUTS converge to a stationary distribution faster and suffer less from autocorrelation within a chain.

2.1.5 Assessing Convergence

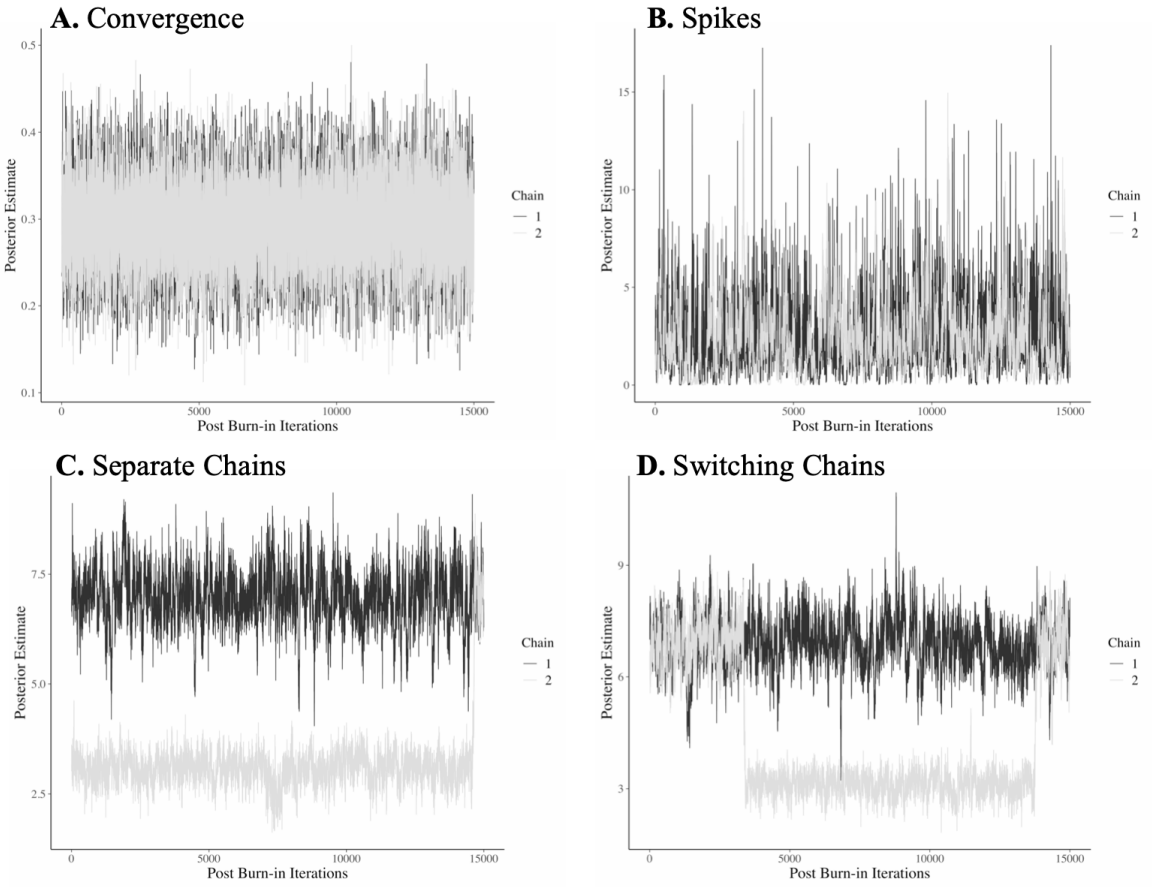


Figure 2. Examples of trace plots showing convergence or non-convergence.

Researchers have exerted considerable effort to develop diagnostics for assessing convergence of MCMC chains, given the importance of reaching stationarity for drawing valid inferences. Convergence needs to be assessed for each parameter in the model. The simplest convergence diagnostic is a visual inspection of the trace plots of the chains. These plots depict the samples that were drawn from the posterior. While the exact value drawn from the posterior changes with each iteration, the trace line(s) should typically show a tight, horizontal band across the plot (Figure 2, panel A). However, for parameters that are naturally skewed (e.g., variance parameters that are constrained to positive values), the trace line(s) may show some spikes on one side of the horizontal band (Figure 2, panel B). Thus, whether these spikes in the trace line(s) indicate non-convergence depends on the type of parameter and whether the spikes reach implausible, extreme areas of the target distribution. The trace plot can also provide evidence of other forms of non-convergence: Two chains may be sampling from different areas of the

target distribution (Figure 2, panel C), or the chains are jumping from one area of the target distribution to another area (Figure 2, panel D). This visual diagnostic is most useful for detecting major issues with convergence and should not be the sole method for assessing convergence. For that reason, several other convergence diagnostics have been developed, such as the Geweke (Geweke, 1992), the Heidelberger and Welch (Heidelberger & Welch, 1983), the Raftery and Lewis (Raftery & Lewis, 1992), and the Gelman-Rubin (Gelman & Rubin, 1992) convergence diagnostic.

Each diagnostic assesses different aspects of the MCMC chains. The Geweke diagnostic is a diagnostic for a single MCMC chain, and it compares the first 10% of the chain to the last 50% of the chain using a *z*-test to see if they differ significantly (Geweke, 1992). If the *z*-statistic is significant, it suggests that the start of the chain has not converged and the burn-in period should be increased. The Heidelberger and Welch diagnostic uses the Cramer-von-Mises statistic to assess non-stationarity in a single MCMC chain (Heidelberger & Welch, 1983). If evidence of non-stationarity is found, the first 10% of the iterations are removed, and the test is repeated until stationarity is found or until 50% of the chain is removed. If the latter occurs, convergence to stationarity is not reached, and the number of iterations should be increased. The Raftery and Lewis diagnostic helps determine three components of the MCMC chains for each parameter: the burn-in period, the total number of iterations, and the thinning interval (Raftery & Lewis, 1992).[2] These three components are determined for a particular quantile and degree of accuracy for the posterior distribution. As the posterior distribution is often assessed using the posterior mean or median, the 0.5 quantile is often selected for computing the Raftery & Lewis diagnostic. Finally, the Gelman-Rubin diagnostic assesses convergence by comparing multiple Markov chains within one Bayesian analysis (Gelman & Rubin, 1992). It compares the estimated within-chain variance to the between-chain variance for each model parameter. Large differences between these variances indicate issues with convergence. The ratio of the two variances is represented by the potential scale reduction factor (PSRF) or the $\hat{R}$- statistic (S P Brooks & Gelman, 1998; Gelman & Rubin, 1992; Vehtari et al., 2019). If the chains have converged, the PSRF for that parameter is close to 1. If the PSRF is > 1, it indicates that the number of iterations should be increased so that either the between-chain variance decreases or the within-chain variance increases (as it explores the full posterior distribution). Several cutoffs have been suggested for concluding convergence, such as < 1.05 in M*plus* (Asparouhov & Muthén, 2010b). Ideally, a researcher uses a combination of these diagnostics to assess convergence for all parameters in the model of interest.

2.1.6 Posterior Point Estimates, Credible Intervals, and Highest Posterior Densities

Once a Bayesian analysis through MCMC methods is conducted and convergence is confirmed, a researcher can use several values to assess the posterior distributions of the parameters. First, point estimates, such as the mean, median, mode, and variance, can be

[2] A thinning interval can be used to reduce autocorrelation between iterations in the chain. When a thinning interval is used, only every s^{th} iteration of the chain is retained for the purpose of summarizing the posterior distribution.

computed. These estimates are all computed using the conditional distribution of a parameter $\boldsymbol{\theta}$, which is obtained by averaging over the marginal distribution of $\mathbf{y}$.

The posterior point estimates of quantiles can be used to construct something called a 95% (or another percentage) *credible* interval. It is important to note that this interval is not the same as the frequentist 95% *confidence* interval. These intervals have different interpretations because of how the parameters are defined within each framework. Recall that in frequentist estimation, the population parameters are considered fixed, whereas, in Bayesian estimation, population parameters are considered random. Thus, because we assume that a parameter has a parameter distribution, the posterior samples obtained through MCMC can be used to obtain quantiles. These quantiles can be used to find the probability that a parameter lies within a particular interval (Kaplan & Depaoli, 2012). In other words, the probability that a parameter lies in a particular 95% credible interval is .95. This interpretation is completely different from the frequentist perspective, where the probability is either 0 or 1 that a fixed population value lies within a particular 95% confidence interval.

In addition to the credible interval, it is also possible to examine an interval referred to as the *highest posterior density* (HPD) interval (Box & Tiao, 1973; Kaplan & Depaoli, 2012). This is an interval of the posterior distribution where every point inside the interval has a higher density than any point outside the interval. Similar to the credible interval, the HPD interval can contain 95% of the distribution, but other percentages, such as 90%, are also common. In contrast to the credible interval, a 95% HPD interval does not need to start at the .025 quantile and end at the .975 quantile. A 95% credible interval and 95% HPD interval will only match if the posterior distribution is unimodal and symmetric. Thus, the HPD interval is beneficial for posterior distributions that are not unimodal or not symmetric.

2.2 Bayesian Structural Equation Models

The topics described in this section assume a basic understanding of SEM estimation. For further resources regarding SEM, I refer readers to Hoyle (2012), Kaplan (2009), Kline (2015), Lei and Wu (2007), and Tarka (2018). The Bayesian approach was first combined with SEM over twenty years ago (Scheines et al., 1999). This combination is not surprising given that the Bayesian framework has certain features that become especially advantageous for SEM. However, it was not until around 2012 that the use of Bayesian estimation of SEMs in the applied literature really started to increase (van de Schoot et al., 2017). In this section, I will discuss the advantages and drawbacks of Bayesian SEM and introduce two ways in which my dissertation research will address some of these drawbacks.

2.2.1 Advantages of Bayesian SEM

Researchers have several reasons, both theoretical and practical, for turning to Bayesian approaches to SEM (van de Schoot et al., 2017). In this section, I will address the major advantages for using a Bayesian approach to SEM.

The inclusion of prior distributions is at the root of many of the advantages of Bayesian estimation for SEM. From a theoretical standpoint, priors allow updating

knowledge instead of starting each study from scratch and testing null hypotheses (van de Schoot et al., 2014). In the absence of prior knowledge, priors can still be used to incorporate uncertainty about the parameters explicitly. This approach is more in line with the overarching goal of science: moving knowledge forward.

From a practical standpoint, the inclusion of prior distributions has several advantages. For example, computational difficulties that are often encountered with maximum likelihood (ML) estimation, such as Heywood cases, can be prevented through the specification of appropriate priors for the (residual) variances (Lee, 2007). In ML estimation, these problematic parameters are often fixed to zero. More generally, in typical SEM studies that use maximum likelihood estimation, many parameters are fixed at zero to ensure that the model is identified. However, it is implausible that all these parameters are exactly zero in the population. Thus, this source of model error is almost always present in frequentist SEMs (B. O. Muthén & Asparouhov, 2012). In Bayesian SEM, fixed zeroes can be replaced by small-variance priors that are centered at zero, allowing for slight deviations from zero. This approach most accurately reflects the uncertainty about the specified model and prior theory about parameter values (MacCallum et al., 2012). In addition, if the 95% credible interval of a parameter with a small-variance prior does not contain 0, then it provides evidence that the parameter is non-zero in the population (B. O. Muthén & Asparouhov, 2012).

Another advantage of Bayesian estimation for SEM is that it can provide reliable statistical estimates with small sample sizes, provided that appropriate priors are specified (Lee, 2007; McNeish, 2016; Smid, Depaoli, et al., 2019; Smid & Winter, 2020). While researchers generally agree that larger sample sizes are better, there are some situations under which collecting a large sample may be challenging (Smid, McNeish, et al., 2019). Examples include studies that focus on naturally small populations (burn victims who needed mechanical ventilation; van de Schoot et al., 2015), hard to access populations (e.g., incarcerated mothers; Zeman et al., 2018), or study designs that result in financial constraints (e.g., ecological momentary assessment; Schwerdtfeger et al., 2020). Small samples often cause problems for researchers using frequentist estimators such as ML, resulting in non-convergence, inadmissible estimates, or inaccurate estimates (Smid, McNeish, et al., 2019). The inaccuracy of the estimates is in part due to ML's reliance on asymptotic (large sample) theory. Bayesian estimation does not rely on this theory; instead it uses MCMC methods to sample from the posterior distribution directly. However, it is important to note that using Bayesian estimation without thoughtfully specified, informative priors can result in posterior estimates that are *more* biased than their ML counterparts (Depaoli, 2013; McNeish, 2016; Smid & Winter, 2020; van Erp et al., 2018).

Bayesian estimation's use of MCMC methods can also help resolve issues of missing data, nonlinearity, and nonnormality (Lee, 2007). An example of a model in which these characteristics are particularly advantageous is a model that contains mediation effects. In these models, two regression coefficients are multiplied, and the resulting parameter is not normally distributed, resulting in biased standard errors and confidence intervals. As Bayesian estimation makes no assumptions about the specific form of the posterior distribution, the standard error (or posterior standard deviation) can be accurately estimated (Y. Yuan & Mackinnon, 2009). Similarly, Bayesian estimation yields unbiased

credible intervals for reliability estimates in multi-level SEMs, where bootstrapped confidence intervals are not appropriate (Geldhof et al., 2014). Concerning issues surrounding missing data, MCMC methods such as the Gibbs sampler do not distinguish between missing data, latent variables, and parameters. All three are considered unknown and random, and are estimated through a joint posterior distribution conditional on the observed data (Asparouhov & Muthén, 2010b; Gelman et al., 2013).

Finally, using an iterative sampling algorithm like MCMC also results in a posterior distribution of each parameter in their model, which provides more information than a single point estimate found through frequentist estimation. Although researchers using frequentist methods can report a 95% confidence interval, this interval is less straightforward to interpret than the Bayesian credible interval (Kaplan & Depaoli, 2012; van de Schoot et al., 2014). In addition, the posterior distribution is not constrained to follow a normal distribution (or any parametric distribution). In contrast, normality is an assumption underlying most frequentist methods for constructing a 95% confidence interval. The rich information provided by the posterior distribution, combined with the computational advantages of the MCMC methods, provide a convincing case for the use of Bayesian estimation for SEMs. However, several drawbacks of Bayesian SEM have also been voiced. These will be discussed in the next section.

2.2.2 Drawbacks of Bayesian SEM

Before discussing the drawbacks of Bayesian SEM in particular, I want to acknowledge a general philosophical criticism of Bayesian statistics, namely its subjectivity (Press, 2003). The practice of including prior knowledge or beliefs directly in an analysis goes against the idea that scientists should be concerned with objective knowledge (Gelman, 2008). Indeed, the Bayesian perspective on probability is that it does not have an objective status but instead "represents the quantification of our experience of uncertainty" (Kaplan, 2014 p. 284). Even if the concept of including prior knowledge in an analysis is accepted, critics often point out that different individuals (e.g., researchers or content experts) may hold different prior beliefs (Press, 2003). Why should my prior belief be used over others' prior beliefs? Proponents of the Bayesian approach have argued that scientific objectivity can still be attained with objective Bayes (objective or reference priors; Press, 2003; Wagenmakers et al., 2008) or the evidence-based use of subjective Bayes (Kaplan, 2014). In this dissertation, I subscribe to the evidence-based use of subjective Bayes. This approach to Bayesian statistics prescribes the use of historical data (e.g., prior research) over personal belief to inform the prior distributions specified for an analysis. If historical data are not available, diffuse or reference priors should be used (Berger, 2006). Through this method of including prior knowledge, the evidence-based use of subjective Bayes aligns with the argument that objective science needs to refer to specific data that inform the research (Jaynes, 1968; Kaplan, 2014).

I will now continue my discussion on the drawbacks associated with Bayesian estimation of SEMs, which appear to boil down to a simple message: Bayesian SEM requires more effort and thought from the researcher than frequentist estimation. In other words, the drawbacks of Bayesian estimation are not necessarily fatal computational issues that have no solution. Instead, they are components of the estimation process that are more cumbersome for Bayesian estimation than for frequentist estimation. In this

section, I will discuss several components of Bayesian estimation where the researcher's active role is required.

First, while the ability to use small-variance priors to relax exact zero constraints within a model has been touted as an advantage to Bayesian SEM, it has also been repeatedly criticized (MacCallum et al., 2012; Stromeyer et al., 2015). While these critiques are slightly different in their approach, they both emphasize that estimating many more parameters in a model may inflate model fit and reduce generalizability and replicability. These non-zero parameters may be capturing random sampling noise rather than true minor deviations from zero present in the population (Stromeyer et al., 2015). Asparouhov and colleagues (2015) have refuted this claim, stating that model fit would only be inflated if the small-variance priors are relaxed to the point that the model becomes equivalent to a model that freely estimates all parameters. They also argue that the point of including small-variance priors is to evaluate the sources of differences between a hypothesized (clean) model and the (messy) data.

Another possible drawback to Bayesian estimation of SEM is that it requires more effort from the researcher than frequentist approaches to SEM (MacCallum et al., 2012). For example, with ML estimation, convergence is reached once the algorithm finds the maximum to some pre-specified level of precision. This level of precision is often chosen by the software or package creators and does not require any input from the researcher (although some may argue that it should require the researcher's input). In contrast, with Bayesian estimation, the researcher needs to specify the number of chains, the number of burn-in iterations and the number of posterior sampling iterations. Next, the researcher needs to assess, for each parameter in the model (which, for SEM, can be many), whether the MCMC algorithm has converged to a stable estimate, using a variety of the convergence diagnostics discussed above. Software packages can automate some of the decisions surrounding convergence, but the researcher will need to be aware of what these automated decisions are and if they are stringent enough for the purposes of their research. Several guides and checklists have been developed to help researchers navigate these decisions for SEMs (e.g., Depaoli & van de Schoot, 2017; Harindranath & Jacob, 2018; Miočević, 2019; Song & Lee, 2012; van de Schoot et al., 2020, 2021). While these guides are helpful, it is true that researchers play a more active part in the estimation process if they use a Bayesian approach to SEM.

Prior specification is another component of Bayesian estimation where active participation of the researcher is required. To help researchers with this step, software packages such as M*plus* (Asparouhov & Muthén, 2010b) and the R package 'blavaan' (Merkle & Rosseel, 2018) have implemented default priors for SEMs. However, researchers have repeatedly demonstrated that naively relying on default priors can have adverse effects on the posterior estimates of SEMs (Depaoli, 2013; Depaoli & Clifton, 2015; Smid, McNeish, et al., 2019; Smid & Winter, 2020; van Erp et al., 2018). In contrast, specifying thoughtful, informative priors in SEMs results in posterior distributions that more closely approximate the population parameter compared to diffuse priors or frequentist approaches (Depaoli, 2013; McNeish, 2016; Smid, Depaoli, et al., 2019; Zondervan-Zwijnenburg et al., 2018). Researchers might be hesitant to specify informative priors because their prior knowledge may not agree with the information provided by the data (i.e., prior-data disagreement). Indeed, researchers have found that

informative, inaccurate priors can negatively impact the posterior estimates (Depaoli, 2014; Dingjing Shi & Tong, 2017). Researchers may also wonder what information they should use as the basis for their priors. Several guides exist that illustrate the process of prior elicitation (Van de Schoot et al., 2021; Zondervan-Zwijnenburg et al., 2017). Researchers also have some tools at their disposal to examine the impact of their priors on the posterior estimates. Most importantly, each Bayesian analysis should be followed by a sensitivity analysis of the priors, in which alternative priors are examined to fully understand the impact of the priors specified for the original analysis (Depaoli et al., 2017, 2020; B. O. Muthén & Asparouhov, 2012; van Erp et al., 2018). In addition, researchers can get an idea of the impact of their prior specification before they analyze their observed data through prior-predictive checks (Evans & Jang, 2010).

Finally, one major hindrance to the adoption of Bayesian SEM in the mainstream literature was a lack of readily available and interpretable indices for model fit and selection (Levy, 2011). Until recently, Bayesian model fit assessment was limited to absolute fit indices such as the posterior predictive *p*-value (PPP-value; Gelman, Meng, et al., 1996; Meng, 1994) and comparative fit indices such as the Bayesian information criterion (Schwarz, 1978), Deviance information criterion (DIC; Spiegelhalter et al., 2002), Bayes Factor (BF; e.g., Wagenmakers, 2007), leave-one-out cross-validation (LOO; Geisser & Eddy, 1979; Gelfand & Dey D.K., 1994), and widely applicable information criterion (WAIC; Watanabe, 2010). The Bayesian model evaluation toolkit was missing a set of approximate fit indices that could be used to assess fit along a continuum and without the need for alternative models. Fortunately, a series of fit indices was recently developed and implemented (Asparouhov & Muthén, 2019; Garnier-Villarreal & Jorgensen, 2019; Hoofs et al., 2018). It is now possible to assess the approximate fit of a Bayesian SEM with Bayesian versions of the root mean square error of approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980), comparative fit index (CFI; Bentler, 1990), and Tucker-Lewis index (TLI; Tucker & Lewis, 1973). Asparouhov and Muthén (2019) implemented these approximate fit indices in M*plus* together with a new version of the posterior-predictive *p*-value (PPP-value) that is more appropriate when there is missing data. While the performance of these new fit indices still needs to be assessed across a variety of conditions, their development is promising for the future of Bayesian SEM research.

2.2.3 Opportunities for Bayesian SEM

My dissertation focuses on two of the drawbacks discussed in the previous section. In Study 1, I assess the performance of the newly implemented model fit indices in M*plus* across a wide variety of conditions, such as population model, location and severity of model misspecification, sample size, amount and location of missing data, and prior specification. In Study 2, I assess three methods for detecting prior-data disagreement across multiple sample sizes in a commonly used SEM: the latent growth model (LGM). With these two studies, I aim to improve the applicability of Bayesian estimation of SEM for applied researchers.

Chapter 3
Study 1: Performance of Model Fit and Selection Indices for Bayesian Structural Equation Modeling

3.1 Introduction

Several methods for model fit assessment and model selection have been implemented for Bayesian SEM. However, until recently, Bayesian model fit assessment was limited to absolute fit indices such as the posterior predictive *p*-value (PPP-value; Gelman, Meng, et al., 1996; Meng, 1994) and comparative fit indices such as the Bayesian Information Criterion (BIC; Schwarz, 1978), Deviance information criterion (DIC; Spiegelhalter et al., 2002), Bayes Factor (BF; e.g., Wagenmakers, 2007), leave-one-out cross-validation (LOO; Geisser & Eddy, 1979; Gelfand & Dey D.K., 1994), and widely applicable information criterion (WAIC; Watanabe, 2010). The Bayesian model evaluation toolkit was missing a set of approximate fit indices that could be used to assess fit along a continuum and without the need for alternative models.

Fortunately, new model evaluation tools were added to the toolkit by the recent introduction of a series of approximate model fit indices for Bayesian SEM (Asparouhov & Muthén, 2019; Garnier-Villarreal & Jorgensen, 2019; Hoofs et al., 2018; Liang, 2020). In addition, the M*plus* implementation of the PPP-value has been adjusted to properly account for missing data (Asparouhov & Muthén, 2019). The new approximate fit indices have great potential as they detect model misspecification to a similar extent as their maximum likelihood (ML) counterparts (Garnier-Villarreal & Jorgensen, 2019; Hoofs et al., 2018) and perform well when data are missing at random (Asparouhov & Muthén, 2020). Another advantage is that they have been implemented in multiple software packages, such as the R package 'blavaan' (Merkle & Rosseel, 2018) and M*plus* (L. K. Muthén & Muthén, 2017). This has aided the adoption of the new fit indices in the applied literature (e.g., Dwirifqi Kharisma Putra et al., 2019; Hanson et al., 2020; Nakadai et al., 2020; Phipps et al., 2020).

However, the implementation of the fit indices differs across software packages, resulting in conflicting results about their performance across simulation studies. Furthermore, they have been studied only under limited conditions. For example, their performance has not been assessed for models with a mean structure (e.g., latent growth models) or relatively small sample sizes. In addition, an important element that may affect the performance of the new model fit indices is missing data.

The presence of missing data is a factor that plays a role in almost any study in the social sciences that relies on SEM (Bell et al., 2014; Graham, 2009; Lang & Little, 2018; Nicholson et al., 2017). Missing data can arise through different mechanisms. Rubin (1976) defined three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR occurs if the

probability of missing a value is not related to any observed variable or the missing value itself. If data are MAR, it means that the probability of missing a value is related to an observed variable but not to the missing value itself. These two missing data mechanisms are often referred to as ignorable missingness (Graham, 2009; R. J. Little & Rubin, 2002). These mechanisms are ignorable in the sense that, as long as they are taken into account in the analysis, they result in unbiased parameter estimates (Graham, 2009; R. J. Little, 2021). In contrast, nonignorable missingness occurs when data are MNAR. Here, the probability of missing a value is related to the missing value itself. With MNAR, there are no straightforward methods to adjust for the missing values, resulting in biased parameter estimates (Graham, 2009). To prevent biased estimates due to MNAR, one needs to uncover and explicitly model the missingness mechanism (R. J. Little, 2021). In the current study, I focus on the MAR mechanism.

While Asparouhov and Muthén (2020) concluded that the new model fit indices performed well when data were M(C)AR, their conclusions are based on a simulation study that generated data from a simple linear regression with one observed outcome and one observed predictor variable. From the frequentist literature, we know that the impact of missing data on model fit indices may depend on many factors, such as the sample size (Enders & Mansolf, 2018), the amount and location of missing data (Wu & West, 2010), and the number of different missing data patterns present in the sample (Zhang & Savalei, 2020). Thus, the focus of the current study is to add to the literature on model fit and selection indices for Bayesian SEM through an extensive simulation study that examines the performance of the PPP-value, BIC, DIC, and approximate fit indices.

The remainder of this section will be organized as follows. First, I will introduce the concept of model misspecification, after which I will present the model fit and selection indices currently implemented in M*plus*. This section is followed by a discussion of the existing literature on factors affecting the ability of model fit and selection indices to detect model misspecification. This discussion will cover both Bayesian and frequentist implementation of model fit indices, as the existing literature on the Bayesian approximate fit indices is still limited. Next, I will introduce and specify the models for which the model fit and selection indices' performance will be examined through a simulation design.

3.1.1 What is Model Misspecification?

SEMs are used to represent the hypothesized theory underlying the data-generating process. For that reason, assessing the model's fit to the observed data is an important step to evaluate whether the theory adequately explains the data-generating process. The adequacy of a specified model can be assessed with two different questions in mind (West, Taylor, & Wu, 2012). First, we can ask: Does the hypothesized model provide a good fit to the observed data? Second, we can ask: If multiple competing models exist, which of these models best represents the observed data? These two questions map on to model fit and model selection, respectively.

Model fit and selection indices all assume that the estimated model is the correctly specified model for the population. However, it is generally accepted that most estimated models are not 100% correct and include some form of *mis*specification (West et al., 2012). Thus, model fit and selection indices should capture the degree to which a model

is misspecified. The definition of model misspecification needs to be clear before this question can be assessed through simulation research. In the literature, model misspecification has been defined using different characteristics. For example, a model can be misspecified if it is either under- or over-parameterized (Hu & Bentler, 1998). Model under-parameterization occurs when one or more parameters are fixed to zero even though their population values are non-zero. Model over-parameterization occurs when one or more parameters are estimated whose population values equal zero. Under- and over-parameterization can also co-occur in different parts of the model.

A related way of looking at model misspecification is by differentiating between internal and external misspecification (Kaplan, 1989, 1990). Internal model misspecification assumes that the correct variables are included in the model but that the estimated associations between these variables are incorrect. In contrast, external model misspecification arises when a key variable of the theoretical model is not included in the statistical model. There might be several reasons why such a variable is excluded. Perhaps it was not included in an original dataset that is now used for secondary analysis, or its importance was not realized until after data collection was completed. Alternatively, the variable could reflect sensitive information, information from an unavailable informant, or hard to access information (Harring et al., 2017).

Another way of categorizing model misspecification is as substantively relevant or irrelevant (Saris et al., 2009a). Substantively relevant misspecifications are misspecifications that lead to incorrect conclusions. For example, in a model that examines whether *Y1* predicts *Y2*, all variables that could cause spurious relationships between the two variables (i.e., third variables) need to be included. If some third variable is not included, the residuals of *Y1* and *Y2* covary. If this covariance is not included in the estimated model, the path from *Y1* to *Y2* may be under- or over-estimated, leading to the wrong conclusion about their association.

In contrast, substantively irrelevant misspecifications are misspecifications that result in a model that is adequate for all practical purposes even though it is not entirely correct. For example, suppose the correlation between two latent factors is .95. In that case, researchers might conclude that this two-factor model can be reduced to a one-factor model (essentially fixing the correlation between the two factors to 1).

Model misspecification in SEM can also be described according to the location of misspecification. While some SEMs, such as path models, can only have misspecification in the covariance matrix, other models include multiple locations where misspecification can arise. For example, in growth curve models, misspecification can exist in the marginal mean structure, conditional mean structure, between-individual covariance structure, and within-individual covariance structure (Wu et al., 2009). Here, the marginal mean structure refers to the average growth trajectory estimates, while the conditional mean structure refers to the individual growth trajectories. The between-individual covariance structure refers to the specification of the variances and covariances among the growth parameters. The within-individual covariance structure refers to the variances and covariances between the observed variables' residual variances. Model misspecification in the marginal mean structure affects the fit of the within- and between-person covariance structures, and under realistic conditions, misspecification of the covariance structures also affects the marginal mean structure (Wu et al., 2009). Further,

the fit of the conditional mean structure is affected by both the marginal mean structure and the covariance structures (Wu et al., 2009). In general, global fit indices used in SEM (including the approximate indices discussed below) can only detect model misfit in the covariance structures and the marginal mean structure (Wu et al., 2009).

In practice, many of the studies that examine the ability of model fit indices to detect model misspecification focus on under-parameterized models (e.g., Fan & Sivo, 2005, 2007; Heene et al., 2012; Hsu et al., 2015; Hu & Bentler, 1998; Leite & Stapleton, 2006; Lin et al., 2017; Mahler, 2011; Maydeu-Olivares et al., 2017; Maydeu-Olivares, 2017; Ryu & West, 2009; Saris et al., 2009b; Savalei, 2012; Dexin Shi et al., 2018; Wu & West, 2010). Examples of under-specification include (a) fixing one or more covariances between factors in a confirmatory factor analysis (CFA) to zero (Fan & Sivo, 2005; Hsu et al., 2015; Hu & Bentler, 1998), (b) fixing one or more non-zero cross-loadings to zero in a CFA (Savalei, 2012; Dexin Shi et al., 2018), or (c) fixing the mean and variance of a quadratic slope in a latent growth model (LGM) to zero (Wu & West, 2010). These examples illustrate that some of the definitions of misspecification overlap. Fixing a covariance between two factors to zero can also represent a substantively irrelevant misspecification if the covariance is close to zero in the population model. Fixing the mean of a quadratic slope to zero results in a misspecification in the marginal mean structure, whereas fixing the variance of a quadratic slope to zero results in a misspecification in the between-individual covariance structure. I will further discuss these nuances after presenting an overview of the model fit and selection indices that are the focus of the current study.

3.1.2 Overview of Model Fit and Selection Indices

In this section, I will introduce the model fit and selection indices that are implemented in M*plus*. When appropriate, I will also mention alternative model fit and selection indices available in other software packages.

3.1.2.1 Information Criteria

The first two Bayesian model selection indices, the BIC (Schwarz, 1978) and the DIC (Spiegelhalter et al., 2002), are absolute fit indices because they do not require a baseline or null model to be computed. They can also be considered comparative fit indices because the BIC and DIC values have no intrinsic meaning and can be interpreted only when they are computed and compared for multiple competing models. Thus, these indices can be used to answer the question, "If multiple competing models exist, which of these models best represents the observed data?" While the BIC and DIC are just two possible information criteria, they are the only information criteria implemented for Bayesian estimation in M*plus*. Some other information criteria are the Widely Available Information Criterion (Watanabe, 2010) and Leave-One-Out information criterion (LOOIC; Geisser & Eddy, 1979).

The BIC and DIC are based on a deviance term that can be calculated using the following equation:

$$Deviance = -2\log\left[p\left(y\middle|\hat{\theta}_{EAP}\right)\right], \quad (10)$$

where $\hat{\theta}_{EAP}$ is the posterior mean estimate, and $\log[p(y|\hat{\theta}_{EAP})]$ is the log likelihood based on that posterior mean estimate.

The BIC (Schwarz, 1978) uses the deviance term defined in Equation (10) as follows:

$$BIC = -2\log[p(y|\hat{\theta}_{EAP})] + p(\log[n]), \tag{11}$$

where p is the number of parameters in the model and n is the sample size. A model with a lower BIC value should be selected over a model with a higher BIC value.

The DIC (Spiegelhalter et al., 2002) uses the deviance term defined in Equation (10) as follows:

$$DIC = -2\log[p(y|\hat{\theta}_{EAP})] + 2p_{DIC}, \tag{12}$$

where p_{DIC} is a model complexity penalty that is computed as

$$p_{DIC} = 2\left(\log[p(y|\hat{\theta}_{EAP})] - E_{post}(\log[p(y|\theta)])\right). \tag{13}$$

In Equation (12), $E_{post}(\log[p(y|\theta)])$ is the posterior mean of the log likelihood, which is computed with the following equation:

$$E_{post}(\log[p(y|\theta)]) = \frac{1}{S}\sum_{s=1}^{S}\log[p(y|\theta^s)], \tag{14}$$

where S is the number of MCMC iterations and θ^s is the posterior sample for parameter θ at the sth draw. Thus, the DIC is only partially Bayesian, as the deviance term and the first term in the computation of p_{DIC} are both based on $\hat{\theta}_{EAP}$, a point estimate of the posterior mean. Only the second term in the computation of p_{DIC} is based on the entire posterior sampling chain. A model with a lower DIC value should be selected over a model with a higher DIC value.

3.1.2.2 Posterior Predictive *p*-value

The PPP-value reflects the extent to which replicated data generated under the posterior estimates of the model are similar to the observed data (Gelman et al., 2013). To assess the similarity of the replicated and observed data, the posterior model-implied covariance matrix for the variables is compared to the data covariance matrix of the replicated and the observed data, using a discrepancy statistic, such as the likelihood ratio test (LRT). This is done at each iteration of the MCMC chain. The PPP-value represents the proportion of posterior predictive discrepancy statistics (computed for replicated data) that are greater than the discrepancy statistics of the observed data. A model that fits the data well is expected to have a PPP-value close to 0.5 (i.e., half of the replicated datasets had greater discrepancy values compared to the observed data). A misspecified model is expected to have a PPP-value close to 0 (i.e., most of the replicated datasets had greater discrepancy values compared to the observed data). Generally, a PPP-value > .05 is considered to indicate a well-fitting model. Thus, this index can be used to answer the

question, "Does the hypothesized model provide a good fit to the observed data?" Some advantages of the PPP-value compared to indices such as the DIC are that it incorporates uncertainty of estimation in the model by using the full posterior distribution, and that it does not depend on asymptotic arguments (Gelman, 2013; Levy, 2011).

The PPP-value implemented in M*plus* is based on a discrepancy function *D*, following standard posterior predictive checking methodology (Gelman et al., 2013). The discrepancy function is the LRT function comparing the estimated model (the H0 model) and the unconstrained mean and variance-covariance matrix model (the H1 model). Starting with M*plus* version 8.4, this discrepancy function is defined as follows:

$$D(Y, \mu_1, \Sigma_1, \mu_0, \Sigma_0) = \mathcal{L}(Y|\mu_1, \Sigma_1) - \mathcal{L}(Y|\mu_0, \Sigma_0), \tag{15}$$

where *Y* represents the data and $\mathcal{L}\left(Y\middle|\mu_j, \Sigma_j\right)$ represents the log-likelihood of Y based on the multivariate normal distribution with mean μ_j and covariance matrix Σ_j. This discrepancy function is computed for the observed data, Y^{obs}, and the replicated data generated during the *i*-th MCMC iteration, Y_i^{rep}. For the observed data, this function is computed as

$$D_i^{obs} = D(Y^{obs}, \mu_{1i}(Y^{obs}), \Sigma_{1i}(Y^{obs}), \mu_{0i}, \Sigma_{0i}), \tag{16}$$

where $\mu_{1i}(Y^{obs})$ and $\Sigma_{1i}(Y^{obs})$ are a random draw of the H1 model parameter estimates for Y^{obs}, and μ_{0i} and Σ_{0i} are the H0 model implied mean and covariance matrix obtained from the *i*-th iteration of the H0 model. Likewise, the discrepancy function for the replicated data is computed as

$$D_i^{rep} = D\left(Y_i^{rep}, \mu_{1i}\left(Y_i^{rep}\right), \Sigma_{1i}\left(Y_i^{rep}\right), \mu_{0i}, \Sigma_{0i}\right), \tag{17}$$

where $\mu_{1i}\left(Y_i^{rep}\right)$ and $\Sigma_{1i}\left(Y_i^{rep}\right)$ are a random draw of the H1 model parameter estimates for Y_i^{rep}. These two values are then used to compute the PPP-value as follows:

$$PPP = P(D^{obs} < D^{rep}) \approx \frac{1}{L}\sum_{i=1}^{L}\delta_i \tag{18}$$

where *L* is the number of iterations in the MCMC chain, and $\delta_i = 1$ if $D_i^{obs} < D_i^{rep}$ and 0 otherwise.

3.1.2.2.1 Computing the PPP-value with missing data

In M*plus* version 8.4, each discrepancy function is defined as a test of the fit function for the observed (i.e., not missing) data only. Thus, if there are missing values, these remain missing in the computation of the discrepancy functions. Similarly, the replicated data at each iteration of the MCMC chain also include missing values if the original observed data include missing values. Every missing value in the observed data is matched with a missing value in the replicated data. So, the observed and replicated data have the same pattern of missing data. The discrepancy function for the observed and replicated data is computed in exactly the same manner. At each iteration in the MCMC chain, the

discrepancy function uses the LRT function for the H0 model with the current H0 model parameters based on the observed data and the H1 model estimates obtained from the incomplete observed and replicated data.

The H1 model estimates of the observed and replicated data are both based on the last draw of a 10-iteration MCMC chain of the H1 model. Asparouhov and Muthén (2020) acknowledge that running such a low number of iterations is not perfect as full convergence is not enforced. However, they argue that it is equitable since the approach is used for the observed and the replicated data, and it is the only way that the speed of computation is not compromised. In addition, the H1 model is fast and simple to estimate since it is the unconstrained model. Moreover, by estimating ten iterations of an MCMC chain, this approach results in an approximation of the H1 model parameter distributions for both the observed and replicated data, which can be used to compute the discrepancy functions. To further ensure that the discrepancy functions of the observed and replicated data can be compared, the same starting values are used in the estimation of each H1 model.

The reason why the missing values remain missing for the observed and replicated data is that it ensures that the replicated data are comparable to the real data under the null hypothesis that the H0 model is correct (Asparouhov & Muthén, 2020). As the missing data mechanism is unknown and not explicitly estimated, it cannot be used to generate missing values in the replicated data. However, as the model estimating assumption for the real data set is assumed to be MAR, using the location of the missing values in the real data to generate missing values in the replicated data can also be considered MAR (Asparouhov & Muthén, 2020). When data are MAR, the likelihood of the observed data is independent of the missing data mechanism (R. J. A. Little & Rubin, 1989). Thus, even though the missing data mechanisms of the observed and replicated data are not identical, the discrepancy functions (which are based on the likelihood) for these two data sets are comparable.

Asparouhov and Muthén (2020) demonstrated that this new approach to computing the PPP-value improves power to detect model misspecification compared to the original approach and appears to work well with missing data when estimating a simple regression model.

3.1.2.3 Bayesian Approximate Fit Indices

In this section, I will discuss a series of Bayesian approximate fit indices that are based on the components of the PPP-value discussed above. Approximate fit indices do not assess the significance of some value but are continuous measures of model-data correspondence (Kline, 2015). Within this group of fit indices, the RMSEA (Steiger, 1990; Steiger & Lind, 1980) is considered an absolute fit index, as it relies only on the extent to which the hypothesized model represents the observed data. Thus, this index answers the question, “Does the hypothesized model provide a good fit to the observed data?”. In contrast, the CFI (Bentler, 1990) and TLI (Tucker & Lewis, 1973) are often called comparative or incremental fit indices because they represent the relative improvement in model fit of the specified model over that of a baseline model. The baseline model is often the independence (null) model, which assumes the covariances between variables are zero (Kline, 2015). The specific baseline model depends on the

software program. In M*plus*, the covariances between the endogenous variables are assumed to be zero, but the covariances between the exogenous variables are estimated. This baseline model is often implausible and a poor fit to the observed data. So, approximate fit indices can be used to answer a question that lies in the middle of the questions for model fit and model selection: "To what extent does the hypothesized model represent the observed data compared to the worst possible model?" (Miles & Shevlin, 2007).

The implementation of the approximate model fit indices in M*plus* (Asparouhov & Muthén, 2020) is based on the work by Garnier-Villarreal and Jorgensen (2019). The Bayesian approximate fit indices follow the population formulas of the frequentist approximate fit indices. In M*plus*, each approximate fit index is computed at each MCMC iteration *s*. Thus, one advantage of Bayesian approximate fit indices is that a posterior distribution is formed for each index. This posterior can be interpreted in the same rich way that posteriors for model parameters are interpreted within the Bayesian framework (e.g., by computing the posterior median, mean, or credible interval).[3] The Bayesian RMSEA (BRMSEA) is computed as follows:

$$BRMSEA_s = \sqrt{max\left(0, \frac{D_s^{obs} - p^*}{(p^* - p_{DIC})N}\right)}\sqrt{G}. \tag{19}$$

In that formula, *s* is the *s*th iteration of the MCMC chain, D_s^{obs} is the discrepancy function for the observed data at the *s*th iteration, *G* is the number of groups in the model, and *N* is the sample size. Further, p^* is the number of parameters in the H1 model (unconstrained model), which is based on the number of groups *G*, the number of dependent variables *p*, and the number of covariates *q* and is computed as follows: $p^* = G(p(p+3)/2 + pq)$. Finally, p_{DIC} is the model complexity penalty term that is also used in the computation of the DIC. This term is also known to reflect the number of *effective* parameters in the H0 model. This value will be close to the actual number of parameters in the H0 model when diffuse priors are used for the model parameters. If informative priors are used, p_{DIC} will be smaller than the actual number of parameters in the H0 model. This discrepancy makes sense if you consider that a small-variance prior centered around zero constrains a parameter to be approximately zero, which is nearly equivalent to a parameter that is fixed to zero (Asparouhov et al., 2015; Hoofs et al., 2018). It should be noted that other implementations of the BRMSEA may instead use p_{WAIC} or p_{LOO} to represent the number of effective parameters (Garnier-Villarreal & Jorgensen, 2019).

The Bayesian CFI and TLI are computed in a similar manner. However, as noted above, these two indices require the specification of a baseline or null model to compare the H0 and H1 models to. Generally, the baseline model is a model in which all variances are estimated but all covariances are fixed to zero. The Bayesian CFI is computed as follows:

[3] In M*plus*, the posterior median and a 90% credible interval are returned for each index. In the current version, the full posterior distribution cannot be extracted.

$$BCFI_s = 1 - \frac{D_s^{obs} - p^*}{D_{B,s}^{obs} - p^*}, \quad (20)$$

where $D_{B,s}^{obs}$ is the baseline model discrepancy function for the observed data computed at the *s*th iteration of the baseline model MCMC estimation. Similarly, the Bayesian TLI is computed as follows:

$$BTLI_s = \frac{\left(D_{B,s}^{obs} - p_{DIC[B]}\right)/\left(p^* - p_{DIC[B]}\right) - \left(D_s^{obs} - p_{DIC}\right)/\left(p^* - p_{DIC}\right)}{\left(D_{B,s}^{obs} - p_{DIC[B]}\right)/\left(p^* - p_{DIC[B]}\right) - 1}, \quad (21)$$

where $p_{DIC[B]} = 2pG$, which means that, for the baseline model, the estimated number of parameters are replaced by the actual number of parameters. As the baseline model is estimated with diffuse priors, the difference between the estimated and actual number of parameters is expected to be inconsequential (Asparouhov & Muthén, 2020).

One reason why the BTLI and BCFI were not introduced in earlier versions of M*plus* is how missing data used to be treated in computing the discrepancy function for the observed (and replicated) data (Asparouhov & Muthén, 2020). In previous versions of M*plus*, the discrepancy function was based on the data after imputing missing values. This approach was problematic for the PPP-value because it essentially used the imputed values at the *s*th iteration to create the replicated data. That approach weakens the power to detect model misspecification because the imputed data are based on the H0 model and thus fit the specified model perfectly, even if the estimated model does not match the population model. For the BCFI and BTLI, including imputed values resulted in an opposite problem. As both the BCFI and BTLI rely on a baseline model that assumes zero correlations between variables, imputed data based on this model will be considerably different from the observed data. Using the imputed data in the computation of the discrepancy function would distort $D_{B,s}^{obs}$ much more than D_s^{obs}, resulting in BCFI and BTLI estimates that are meaningless and offer no valuable information about the fit of the H0 model. With the current method for computing the discrepancy functions, these problems are resolved as the method does not include imputed data in its computation.

3.1.3 Factors Impacting Model Fit and Model Selection Assessment

In an ideal world, the model fit and selection indices presented above would reflect only the extent to which a model is misspecified. In reality, several nuisance factors impact the indices, sometimes in surprising ways. Here, I focus on four commonly considered nuisance factors: sample size, location of misspecification, missing data, and the role of priors. For each factor, I will discuss the literature on the Bayesian model fit and selection indices. As this literature is still limited in scope for the approximate and absolute fit indices, BRMSEA, BCFI, and BTLI, I will supplement this discussion with findings from the frequentist literature on model fit and selection.

3.1.3.1 Sample Size

Every index of model fit or selection is affected by sample size. However, the impact of sample size differs across indices and is more straightforward for some than for others.

For example, the DIC's ability to select the correctly specified model generally improves as the sample size increases (Cain & Zhang, 2019; Liang & Luo, 2019; Lu et al., 2017; Zhu & Stone, 2012). The DIC may be more sensitive to detecting model misspecification than the PPP-value (Liang & Luo, 2019). However, the DIC is not a consistent measure of model fit, so we cannot assume that it will increasingly select the true model out of a fixed set of models with increasing sample size (Spiegelhalter et al., 2002, 2014). To draw appropriate conclusions regarding model selection, it may be crucial to look at the magnitude of a difference in DIC between two models instead of assessing if there is any difference at all. Cain and Zhang (2019) found that for smaller samples ($n < 150$), the difference between two models' DIC values needed to be at least 7 to minimize the false selection rate of misspecified models. This threshold could be lowered to 3 for models with larger samples ($n \geq 250$).

Similar to the DIC, the PPP-value becomes increasingly sensitive to model misspecification with increasing sample size (Asparouhov & Muthén, 2010a, 2020; Cain & Zhang, 2019). In addition, the PPP-value is unlikely to reject a correctly specified model regardless of the sample size (Asparouhov & Muthén, 2010a). However, compared to the frequentist LRT test for model fit, the PPP-value is generally less likely to reject a misspecified model across a range of sample sizes. This lower sensitivity may be due to the somewhat arbitrary nature of the .05 cutoff of the PPP-value for rejecting a model. In contrast to the frequentist *p*-value, the distribution of the PPP-value is not uniform between 0 and 1 if the H0 model is true. Instead, its distribution is not known and depends on the specific model estimated. However, a cutoff value of .05 is still used in the way that the H0 model (the estimated model) is rejected if the PPP-value is smaller than .05. As the distribution of the PPP-value is not known, this cutoff value does not necessarily represent the fifth percentile of the distribution. For some models, the fifth percentile is associated with a slightly larger PPP-value. Thus, using the cutoff of .05 may reduce power to detect model misspecification. However, it also lowers the probability of a Type I error, which may be desirable in certain contexts. Asparouhov and Muthén (2020) suggest a method for adjusting the cutoff value that may be useful when the PPP-value is between .05 and .25. In this approach, the posterior estimates of the H0 model are used to create a population model for a simulation. The output of this simulation is then used to generate a distribution of the PPP-value under the assumption that the H0 model was the true model. This distribution can then be used to find the fifth percentile for the specific H0 model in question. In an example, Asparouhov and Muthén showed that this method increased the power to detect model misspecification from .62 to .87.

A more complicated picture emerges for the approximate fit indices. To start, Asparouhov and Muthén (2020) strongly urge against the use of the new Bayesian approximate fit indices in the context of small sample sizes ($n < 200$). For small samples, the PPP-value typically rejects the model (i.e., is < .05) only when the model is severely misspecified. In those cases, the approximate fit indices should be ignored. Further, if the PPP-value does not reject the model (i.e., is ≥ .05), then the approximate fit indices are not needed to provide further evidence of model fit (Asparouhov & Muthén, 2020). These recommendations call into question the utility of the approximate fit indices for sample sizes common in social science research. However, they are in line with

conclusions from the frequentist literature on approximate fit indices, which are likely to reject correctly specified models when sample sizes are small (Heene et al., 2012; Sharma et al., 2005). In contrast, whereas frequentist approximate fit indices do not necessarily become more accurate for larger sample sizes (Ainur et al., 2017; Heene et al., 2012; Leite & Stapleton, 2006; Sharma et al., 2005), the Bayesian approximate fit indices stabilize and become more accurate as the sample size increases (Asparouhov & Muthén, 2020; Garnier-Villarreal & Jorgensen, 2019).

Furthermore, the Bayesian approximate fit indices benefit from the fact that their posterior distribution is available. Thus, a credible interval can be extracted and used to assess if fit is good, inconclusive, or poor using some pre-selected cutoff-value. While the RMSEA, CFI, and TLI were not developed for making binary decisions about model fit, the use of several suggested cutoff values for concluding that a model fits the data well (e.g., CFI/TLI $\geq .90$ or .95; RMSEA $\leq .05$ or .06) is widespread in the applied literature. Their implementation in the frequentist framework has been criticized (e.g., Leite & Stapleton, 2006; Mcneish & Hancock, 2018; Niemand & Mai, 2018; Xia & Yang, 2018), as the cutoff-values are based on very specific population models and conditions, such as a moderately large sample size. For example, the cutoff-values for the TLI and RMSEA are likely to reject correctly specified models when the sample size is relatively small (Sharma et al., 2005). Perhaps more problematic, as the sample size increases, the cutoff-values of the RMSEA (and to a lesser extent the TLI) become more likely to fail to reject a misspecified model (Sharma et al., 2005).

Using the posterior credible intervals through the Bayesian framework can provide a level of nuance to the use of cutoff values. For example, if a 90% posterior credible interval of the BCFI is below .95, we can conclude with 90% certainty that the model is a poor fit to the data. Using the credible interval approach also allows for an inconclusive conclusion: if the 90% credible interval contains .95, we cannot be sure if the model fits the data well or not. That approach may be particularly beneficial for smaller samples. Indeed, Asparouhov and Muthén (2020), demonstrated that using rejection-rates based on the approximate fit indices' credible intervals resulted in more accurate conclusions for small sample sizes ($n = 100$) compared to using the point-estimates. For such small sample sizes, approximate fit indices are more variable across samples (Fan et al., 1999; Sharma et al., 2005). Thus, an analysis based on a small sample is more likely to imply poor (or great) model fit independent from the actual presence or absence of model misspecification. With small sample sizes, the posterior distribution of the approximate fit indices may also be wider, resulting in a credible interval that is less precise. A wider credible interval may be beneficial in the sense that the conclusion will more often be "inconclusive", instead of incorrectly concluding good or poor model fit. It is important to note that Garnier-Villarreal and Jorgensen (2019) do not recommend using fixed cutoff values for the BCFI and BTLI because the cutoffs vary across several study and model characteristics. In addition, they argue that existing cutoff recommendations for the RMSEA based on its frequentist confidence interval likely need to be adjusted for the BRMSEA, as its credible interval is often narrower.

3.1.3.2 Location of Misspecification

As stated above, the location of misspecification (in the covariance structures or the marginal mean structure) may influence model fit and selection indices. Research on this factor is limited, particularly in the Bayesian literature. One reason for this scarcity of knowledge may be that the common choice of population model in this field of research (i.e., a CFA) does not include the marginal mean structure.

One study examined several approaches for computing the PPP-value to detect model misspecification in different structures of LGMs (Fay et al., 2020). The LRT-based version implemented in M*plus* emerged as the superior option among the choices examined. This PPP-value detected misspecification in the covariance structures and both mean structures (both separately and combined). As frequentist approximate fit indices are not sensitive to misspecification in the conditional mean structure (Wu et al., 2009; Wu & West, 2010, 2013), this finding points towards an area in which the PPP-value may be particularly beneficial for assessing model fit. Although the Bayesian approximate fit indices have not been tested for LGMs, we can draw from research on frequentist approximate fit indices for some insights about their ability to detect misspecification in different model structures. Findings from these studies indicate that the RMSEA, CFI, and TLI are less sensitive to misspecification in the marginal mean structure than to misspecification of within- or between-individual covariance structure (Wu et al., 2009; Wu & West, 2010). Among the three indices, the RMSEA appears less sensitive to the magnitude of inter-individual variability in change over time than the CFI and TLI, which makes it a better fit index for detecting the presence of a quadratic slope (Leite & Stapleton, 2006). Moreover, if misspecification in one location (e.g., the between-person covariance structure) is more severe than in another location (e.g., the marginal mean structure), the RMSEA, CFI, and TLI are less sensitive to the additional misspecification in the latter area (Wu & West, 2010). This decrease in sensitivity may be because misspecification in one model structure is confounded by misspecification in another model structure (K.-H. Yuan et al., 2019). Given that the Bayesian approximate fit indices are based on the same information as the PPP-value, it is unclear whether these indices will be able to detect model misspecification in the conditional mean structure.

3.1.3.3 Missing Data

The presence of missing data is a reality in most studies in the social sciences that rely on SEM (Bell et al., 2014; Graham, 2009; Lang & Little, 2018; Nicholson et al., 2017). For that reason, it is important to understand how missing data affects the performance of model fit and selection indices in Bayesian SEM. Before turning to the potential impact of missing data in the assessment of model fit, it is helpful to discuss how missing data are handled in Bayesian estimation. Bayesian estimation uses a process called data augmentation (DA; Dyk & Meng, 2001; Tanner & Wong, 1987). DA is a technique that is also used in ML estimation through the expectation-maximization (EM) algorithm (Dempster et al., 1977). In the Bayesian framework, DA is implemented through the MCMC method (Gelman et al., 2013). One popular version of data augmentation uses the Gibbs sampler. After a researcher specifies a set of starting values, a DA algorithm alternates between two steps:

Imputation (I) step: Predict values for the missing data based on the model, the current parameter estimates and the observed data. The result is a complete data set that includes the observed and the imputed data.

Posterior (P) step: Using the complete data from the I step, the specified prior distributions, and the model, draw a new set of parameters for the mean vector and covariance matrix.

A Markov chain is formed by repeating those two steps many times. In case of the Gibbs sampler, the parameter vector θ is divided into *d* sub-vectors, $\theta = (\theta_1, \ldots, \theta_d)$. Within each iteration, the Gibbs sampler goes through all *d* sub-vectors and draws new values from their posterior distributions conditional on the latest values of the other sub-vectors of θ (Gelman et al., 2013). In M*plus*, the Gibbs sampler moves through three blocks: parameters, latent variables, and missing observations (Asparouhov & Muthén, 2010b). These augmentation methods work well when missing data are ignorable (M[C]AR) and can incorporate a model for the missing-data mechanism if data are MNAR (Gelman et al., 2013).

With missing data, the DIC can be constructed in different ways, and its use and interpretation vary (Celeux et al., 2006). This variety illustrates a more general issue with the DIC, in that it is not based on a universal principle that makes it generically applicable while remaining computationally feasible (Spiegelhalter et al., 2014). The presence of a large number of missing values inflates the effective number of parameters, p_{DIC}, and increases the variability of the DIC estimate (Celeux et al., 2006). Thus, using the DIC for model selection in the presence of missing data may not be appropriate.

With regard to the PPP-value, I will limit the discussion to the version that is currently implemented in M*plus* and was described in Section 3.1.2.2 (Asparouhov & Muthén, 2020). As this new version of the PPP-value treats missing data in a substantially different way compared to previous implementations of the PPP-value, it would not be meaningful to look at research based on these older implementations. Asparouhov and Muthén (2020) examine the impact of varying amounts (25 or 50%) of MCAR data for a two-factor CFA and a fixed amount (50%) of MAR data for a simple linear regression model.

Across these two simulation studies, the PPP-value's ability to detect model misspecification was better when samples were large (i.e., $n = 1000$) or when the amount of missing data was relatively small (25%). Notably, the PPP-value was unlikely to reject the true model across sample size levels (i.e., $n = 300$ or 1000) and missing data rates, retaining acceptable Type I error rates. Their results indicate that the type of missing data (MCAR or MAR) does not affect the performance of the PPP-value.

Similarly, for the Bayesian approximate fit indices, I will also rely on the study performed by Asparouhov and Muthén (2020), as their implementation is connected to the implementation of the PPP-value. For MCAR data in a three-factor CFA, they found that the Bayesian approximate fit indices were affected by sample size and the amount of missing data. The indices became increasingly likely to indicate good model fit for misspecified models in the presence of missing data (10%, 25%, or 50%) and for small sample sizes ($n \leq 300$). The authors argue that approximate fit indices should not be used

for such small sample sizes and demonstrate that the PPP-value is likely to reject the misspecified models under these circumstances. However, for larger sample sizes (e.g., $n \geq 1000$), the approximate fit indices may be informative because, for these larger samples, the PPP-value tends to reject even trivial misspecifications. These findings mimic results from the literature on frequentist multiple imputation for SEM, where the RMSEA, CFI, and TLI also struggle when sample sizes are small (e.g., $n = 100$) and missing data rates are relatively high (e.g., ≥ 30%; Enders & Mansolf, 2018).

3.1.3.4 The Role of Priors

Priors affect the marginal likelihood of a model, and in turn, it is reasonable to assume that priors also affect the model fit and selection indices based on this marginal likelihood (Gelman et al., 2017). Two aspects of priors that may affect model fit and selection indices are the level of informativeness of the prior and the extent to which the prior diverges from the data likelihood.[4] Some research has focused on the impact of the prior specification on model fit and selection indices. So far, this research has mostly looked at the DIC and PPP-value. For the DIC, true model selection rates decrease with increasingly diverging priors (Cain & Zhang, 2019), but the impact of priors lessens with increasing sample size. When priors align with the data likelihood, the DIC may prefer informative priors over diffuse priors, particularly for difficult-to-estimate parameters (Ward, 2008). This preference may be related to the effective number of parameters term used in the computation of the DIC. With diffuse priors, the estimated number of parameters approximates the model's actual number of parameters. With informative priors, the estimated number of parameters will decrease, lowering the DIC's penalty term (Asparouhov & Muthén, 2020). In terms of model selection, increasingly informative priors result in higher true positive rates (Liang & Luo, 2019). However, these informative priors may lead the DIC to become too sensitive for small samples, flagging trivial, or substantively ignorable, misspecifications.

The PPP-value is relatively robust to small variations in the priors (de la Horra & Rodriguez-Bernal, 2003). Similarly, it does not appear to be affected by increasingly informative priors (Liang, 2020). In fact, informative priors that agree with the data likelihood may improve the PPP-values ability to assess model fit (Gelman, Bois, et al., 1996). However, the PPP-value is sensitive to more severely diverging priors, particularly for smaller sample sizes (Cain & Zhang, 2019).

The studies that introduced the Bayesian approximate fit indices implemented in M*plus* only included default or diffuse priors and did not examine any other prior specifications through a simulation study (Asparouhov & Muthén, 2020; Garnier-Villarreal & Jorgensen, 2019). Only one study that examined the Bayesian approximate fit indices, together with the PPP-value and DIC, through a simulation design focused on the prior specification (Liang, 2020). Liang (2020) focused on a particular application of priors, namely, to relax the exact-zero constraints on cross-loadings through small-variance priors. The simulation design also included several diffuse prior conditions for the model's other parameters. Results showed that using small-variance priors that were

[4] While it is tempting to call these priors *inaccurate*, a Bayesian researcher would argue that the data likelihood is just as likely to be the source of inaccuracy. Thus, in this dissertation, I will use *diverging* to describe a disagreement between the prior and the data.

too narrow resulted in inflated rejection rates for the PPP-value but not for the BRMSEA, BCFI, and BTLI (using cutoff-values). In contrast, decreasing the small-variance prior resulted in lower DIC values (this may be because the number of effective parameters decreases as the priors become more informative). For the other model parameters, some diffuse prior specifications (e.g., the M*plus* default prior settings) resulted in slightly inflated rejection rates based on the PPP-value. The DIC and approximate fit indices were not affected by alternative diffuse prior specifications.

So far, the impact of priors on model fit and selection indices has been investigated only in the context of correctly specified models. Thus, it is still unclear how the priors may affect these indices in the context of model misspecification. Diverging, informative priors may mask model misspecification.

3.1.4 Models Examined in this Study

In this study, I explore three different population models: two versions of a three-factor CFA with 5 items per factor, and a 5 time-point LGM with a linear and quadratic slope. I selected those two general population models because they represent two popular SEMs, allowing me to examine misspecification in the covariance (CFA and LGM) and the mean structure (LGM) in models that are relevant because they are in common use. All observed variables are continuous and normally distributed. Each model will be discussed in more detail below.

3.1.4.1 CFA-Simple

The first population model is a three-factor CFA with 5 items per factor that has a simple structure. This means that the model does not include any non-zero cross-loadings (Figure 3).[5] The general CFA can be expressed in the following matrix form:

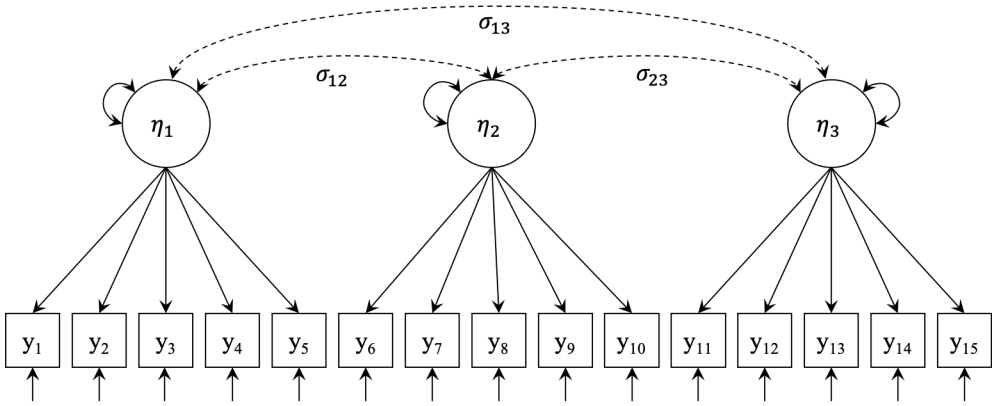
$$\mathbf{Y} = \boldsymbol{\Lambda\eta} + \boldsymbol{\delta}, \tag{22}$$

where $\mathbf{Y}$ represents a vector of observed item responses *i*, $\boldsymbol{\eta}$ represents a vector of latent variables *f*, $\boldsymbol{\Lambda}$ is a loading matrix that related the observed responses to the latent variables, and $\boldsymbol{\delta}$ is a vector of residuals. The covariance matrix of the observed data can be expressed as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\boldsymbol{y}}' + \boldsymbol{\Theta}_{\boldsymbol{\delta}}, \tag{23}$$

where $\boldsymbol{\Phi}$ is the covariance matrix of the latent variables, $\boldsymbol{\Sigma}$ is the population covariance matrix, and $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ is the covariance matrix of residuals. This matrix is diagonal in the absence of residual covariances. Furthermore, the following assumptions are made: $E(\boldsymbol{\eta}) = 0$, $E(\boldsymbol{\delta}) = 0$, and $Cov(\boldsymbol{\eta},\ \boldsymbol{\delta}) = 0$.

[5] The specific design conditions that link to the values in Figure 3, 4, and 5 will be described in the Design section.



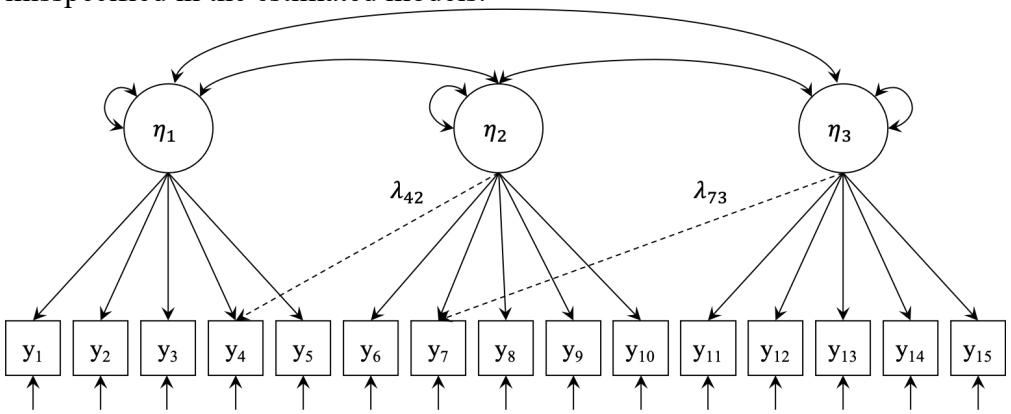
Level 1 misspecified model: $\sigma_{12} = 1$ (i.e., two-factor CFA)
Level 2 misspecified model: $\sigma_{12} = \sigma_{13} = \sigma_{23} = 1$ (i.e., one-factor CFA)

$$\Phi = \begin{bmatrix} 1.00 & & \\ .90 & 1.00 & \\ .90 & .90 & 1.00 \end{bmatrix}$$

$$\Lambda_y^T = \begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

$$diag(\Theta_\delta) = [.51 \quad .51 \quad .44 \quad .36 \quad .36 \quad .51 \quad .51 \quad .44 \quad .36 \quad .36 \quad .51 \quad .51 \quad .44 \quad .36 \quad .36]$$

Figure 3. Path diagram and population parameters for Simple Confirmatory Factor Analysis Model (CFA-Simple). Dotted paths represent population parameters that were misspecified in the estimated models.



Level 1 misspecified model: $\lambda_{42} = 0$
Level 2 misspecified model: $\lambda_{42} = \lambda_{73} = 0$

$$\Phi = \begin{bmatrix} 1.00 & & \\ .50 & 1.00 & \\ .40 & .30 & 1.00 \end{bmatrix}$$

$$\Lambda_y^T = \begin{bmatrix} .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .50 & .00 & .70 & .70 & .75 & .80 & .80 & .00 & .00 & .00 & .00 & .00 \\ .00 & .00 & .00 & .00 & .00 & .00 & .50 & .00 & .00 & .00 & .70 & .70 & .75 & .80 & .80 \end{bmatrix}$$

$$diag(\Theta_\delta) = [.51 \quad .51 \quad .44 \quad .11 \quad .36 \quad .51 \quad .26 \quad .44 \quad .36 \quad .36 \quad .51 \quad .51 \quad .44 \quad .36 \quad .36]$$

Figure 4. Path diagram and population parameters for Complex Confirmatory Factor Analysis Model (CFA-Complex). Dotted paths represent population parameters that were misspecified in the estimated models.

3.1.4.2 CFA-Complex

The second population model follows the same general three-factor CFA set-up as the CFA-Simple model. However, this model has a *complex* structure, and includes two moderate cross-loadings (Figure 4). The two cross-loadings are associated with two items that have their main loading on the first factor.

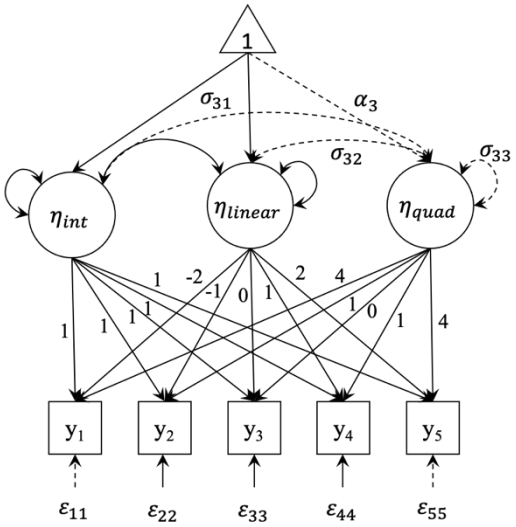
3.1.4.3 LGM

The third population model is an LGM with 5 time points and a quadratic slope (Figure 5). The model can be expressed in the following matrix form:

$$\mathbf{Y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}, \tag{24}$$

where **Y** represents a vector of repeated measures variables, **η** represents a vector of latent variables (the growth parameters), and **Λ** is a fixed loading matrix relating the growth parameters to the observed outcomes. The first column of **Λ** defines the intercept and is a column of ones. Each additional column represents a specific slope (e.g., linear, quadratic). For the population model in Figure 5, these values are -2, -1, 0, 1, 2 for the linear slope, and 4, 1, 0, 1, 4 for the quadratic slope. That means that the intercept is located at the third time point. Finally, $\boldsymbol{\varepsilon}$ represents a vector of residuals. Further, $E(\boldsymbol{\eta}) = \boldsymbol{\alpha}$, a vector of means of the latent variables, **Φ** is the covariance matrix of the latent variables (between-individual covariance matrix), **Σ** is the population covariance matrix, and $\mathbf{\Theta}_{\boldsymbol{\varepsilon}}$ is the covariance matrix of residuals. This matrix is diagonal in the absence of residual covariances. Furthermore, we assume $Cov(\boldsymbol{\eta}, \boldsymbol{\delta}) = 0$. Following this, the covariance matrix of the observed data can be expressed as follows:

$$\mathbf{\Sigma} = \mathbf{\Lambda}_{\boldsymbol{y}}\mathbf{\Phi}\mathbf{\Lambda}_{\boldsymbol{y}}' + \mathbf{\Theta}_{\boldsymbol{\varepsilon}}. \tag{25}$$



Level 1 misspecified model: $\varepsilon_{11} = \varepsilon_{22} = \varepsilon_{33} = \varepsilon_{44} = \varepsilon_{55}$
Level 2 misspecified model: $\sigma_{33} = 0, \sigma_{32} = 0$, and $\sigma_{31} = 0$
Level 3 misspecified model: $\sigma_{33} = 0, \sigma_{32} = 0, \sigma_{31} = 0$, and $\alpha_3 = 0$

$$\Phi = \begin{bmatrix} 3.00 & & \\ 0.24 & 0.30 & \\ 0.09 & 0.03 & 0.04 \end{bmatrix}$$

$$\alpha = \begin{bmatrix} 1.00 \\ 0.30 \\ -0.08 \end{bmatrix}$$

$$diag(\Theta_\varepsilon) = [.20 \quad .50 \quad .50 \quad .50 \quad .80]$$

Figure 5. Path diagram and population parameters for Latent Growth Model (LGM). Dotted paths represent population parameters that were misspecified in the estimated models.

3.1.5 Overview of the Current Study

We need to broaden our knowledge on the performance of the new model fit indices for Bayesian SEM across a large range of conditions. For that reason, the first study of my dissertation examines the performance of the new and updated model fit indices across a set of common SEM models (CFA and LGM) and under a variety of missing data patterns. I looked at the performance of both the point estimates and credible intervals in combination with cutoff values. Other factors that I varied were as follows: sample size, amount and location of missing data, the location and severity of model misspecification, and the impact of the prior specification.

3.2 Design

3.2.1 Population Models

The population values for the CFA-Simple model are included in Figure 3. They are based on a series of previous studies that have used this model to examine the impact of model misspecification (e.g., Garnier-Villarreal & Jorgensen, 2019; Hu & Bentler, 1998). The scale of the latent factors is set by fixing the factor variances to 1 and estimating all factor loadings. This means that the factor loadings are standardized and the residual variance for item *i* loading on factor *f* can be calculated as $1 - \lambda_{if}^2$.

For the CFA-Complex model, the population values for the main factor loadings are the same as for the CFA-Simple model (Figure 4). In that model, the factor covariances reflect moderate associations between the three factors. When there is a non-zero cross-loading, the residual variance of item *i* loading on factors *f* and *g* can be calculated as $1 - (\lambda_{if}^2 + \lambda_{ig}^2)$.

Population values for the LGM population model are included in Figure 5 and are based on a subset of population values examined by Wu and West (2010).

3.2.2 Severity of Misspecification

For each population model, I examined several levels of misspecification. The dashed lines in Figure 3, 4, and 5 represent the paths involved in the misspecification. For the simple and complex CFA model, I followed Garnier-Villarreal and Jorgensen (2019): For the simple CFA model, minor misspecification was introduced by fixing the covariance between the first and second factor to 1, reducing the model to a two-factor CFA (Figure 3). Severe misspecification was introduced by fixing the covariances between all factors to 1, reducing the model to a one-factor CFA. For the complex CFA model (Figure 4): minor misspecification was introduced by fixing one cross-loading to 0, and severe misspecification will be introduced by fixing both cross-loadings to 0. As the current study examines conditions that are different from previous research, using the same misspecification conditions allows for comparison and extension of previous findings.

For the LGM (Figure 5), I examined three distinct types of misspecification, ranging from subtle to more severe: (1) constraining the measurement errors of the observed variables to be equal, (2) fixing the variance of the quadratic slope to zero, and (3) fixing the mean and the variance of the quadratic slope to zero. These conditions are similar to those examined by Wu and West (2010).

Table 1. Population RMSEA values of model misspecification conditions.

Population Model	Misspecification	RMSEA
Simple CFA	1. Three-factor model as two-factor	.039
	2. Three-factor model as one-factor	.053
Complex CFA	1. One cross-loading fixed to 0	.068
	2. Two cross-loadings fixed to 0	.094
LGM	1. Fixed measurement errors	.046
	2. No quadratic slope variance	.071
	3. No quadratic slope	.115

Table 1 shows the population RMSEA values of each model specification to compare the severity of the misspecification across the three population models. These values show that the misspecifications introduced in the Simple CFA model might be considered substantively irrelevant as both RMSEA values are below the commonly used cutoff value of .06. I included these relatively small misspecifications to assess whether the model fit and selection indices may be overly sensitive to minor misspecifications under certain circumstances. In contrast, both levels of misspecification introduced in the Complex CFA model may be considered substantively relevant. Finally, the misspecifications introduced in the LGM cover the largest range of the RMSEA.

3.2.3 Sample Size

Previous research on the performance of model fit indices for SEM has repeatedly identified sample size as an important factor, with larger sample sizes resulting in better performance (Garnier-Villarreal & Jorgensen, 2019; e.g., Heene et al., 2012; Kenny & Mccoach, 2003; Sharma et al., 2005; Dexin Shi et al., 2019). However, Bayesian estimation is often used to increase the available information through the use of priors when only small samples are available (Smid, McNeish, et al., 2019). For that reason, it is important to examine whether the new Bayesian model fit indices can provide some information about model fit even when sample sizes are small. To answer this research question, four different sample sizes were included: $n = 50$, $n = 100$, $n = 250$, and $n = 500$.

3.2.4 Amount of Missing Data

The central aim of the current study is to examine the impact of missing data on Bayesian model fit assessment. Varying the amount of missing data is the most straightforward way of addressing that aim. Previous simulation studies have often included a condition with no missing data (0% missing) and a condition with half of the data missing (50% missing) to reflect a best and worst-case scenario (e.g., Asparouhov & Muthén, 2019; Zhang & Savalei, 2020). I examined a series of systematic reviews on missing data across a variety of fields and found that a missing data rate of 15% is a representative intermediate amount of missing data (Bell et al., 2014; Fiero et al., 2016; Peugh & Enders, 2004; Rioux & Little, 2019). Thus, three different amounts of missing data were included: 0% (best-case), 15% (typical-case), and 50% (worst-case). The way in which these percentages were applied to the observed variables differed slightly across population models. For the CFA models, the percentage reflects the number of missing

values within variables that have missing data (see Section 3.2.4 below). For the LGM, the interpretation of the percentage depends on how many variables have missing data. If missing data are present in one variable, the percentage reflects the number of missing values within that variable. If missing data are present in multiple variables, the percentage reflects the number of missing values across all observed variables that have missing data. The missing data were generated as MAR for all population models. The way in which observed variables are related to the missingness in other variables will be further explained in the Data Generation section (Section 3.2.8).

3.2.5 Number of Variables with Missing Data

Missing data can affect just a few or most variables. Thus, the number of variables with missing data was varied to assess the potential impact of the spread of missing data. The specific number of variables affected by missing data depended on the population model. For the CFA, missing data were present in 1 item per factor (3 items total) or 3 items per factor (9 items total). For the LGM, missing data were present in 1 variable (time point 5) versus 4 variables (time points 2 to 5).

3.2.6 Location of Missing Data (for CFA-Complex)

Missing values can be located in the variables that are involved in the model misspecification or in variables that occur in correctly specified parts of a model. From the frequentist literature, we know that the RMSEA and CFI indicate better model fit when the missing data are located in the part of the model with a misspecification (Zhang & Savalei, 2020). This factor was assessed only for the CFA-Complex population model because it was not straightforward to vary the location of missing data in the CFA-Simple and LGM population models. For the CFA-Simple model, misspecification merges multiple factors to a single-factor model (which includes all observed variables). For the LGM, misspecification affects either the residual variances of all observed variables or the latent quadratic slope effect (which is related to all observed variables). Thus, for the CFA-Complex model, missing data was or was not present in the variables that had cross-loadings.

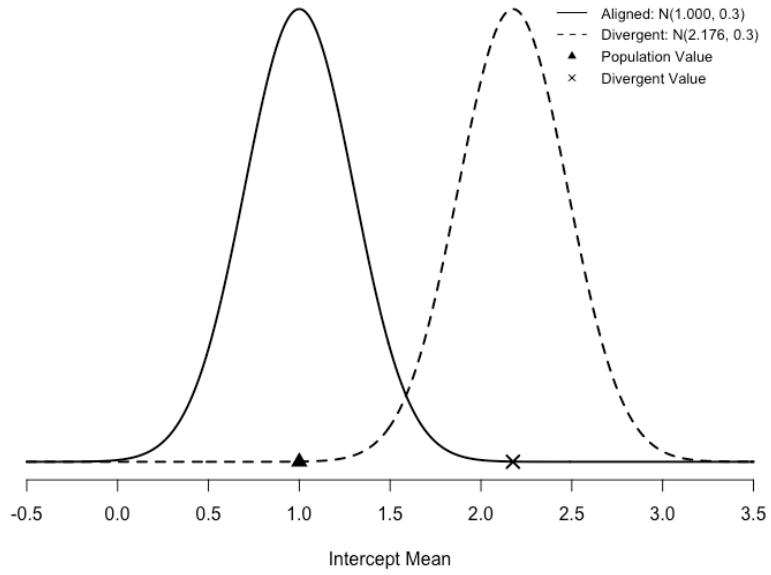
3.2.7 Prior Specification (for LGM)

Several prior specification conditions were included for the LGM population model to investigate the impact of accurate and inaccurate prior distributions on the ability of model fit indices to detect model misspecification. These were: (1) diffuse, (2) narrow and centered over the population value (aligned), and (3) narrow and centered away from the population value (divergent). To limit the number of cells in the simulation, the prior specification condition was only varied for cells with complete data. In the diffuse condition, the M*plus* default priors were used for all parameters.[6] For the aligned and divergent conditions, narrow priors following a normal distribution, $N(\mu, \sigma)$, were placed

[6] The relevant default priors are: N(mean = 0, variance = 10^{10}) for (latent) mean parameters, $IW(\mathbf{I}, p + 1)$ for the (latent) covariance matrix (here, p refers to the number of latent factors and $\mathbf{I}$ is an identity matrix of size p), and IG(shape = -1, scale = 0) for the residual variances (Asparouhov & Muthén, 2010b).

affected by the first two levels of model misspecification (which focus on the covariance structures). For the misspecification in the quadratic slope mean, the population value of

A. Prior Specifications for the intercept mean



B. Prior Specifications for the slope mean

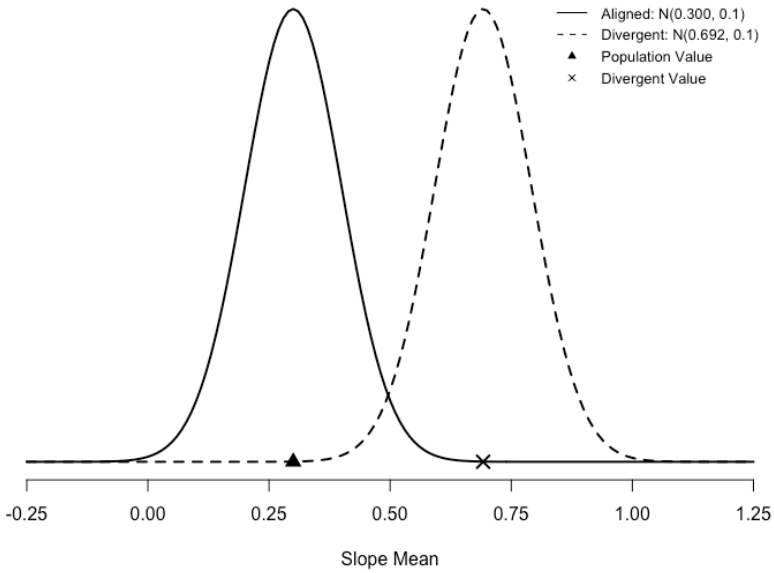


Figure 6. Prior conditions for the intercept mean (panel A) and slope mean (panel B).

the intercept decreased by .115, whereas the linear slope mean increased by .025. Thus, for this condition of model misspecification, the divergent priors are slightly more on the intercept and linear slope means. I selected a standard deviation of 0.3 for the intercept mean and 0.1 for the slope mean.[7]

For the aligned condition, priors were centered on the population values (1.0 and 0.3, respectively). For the divergent condition, priors were centered on a population value such that the resulting prior had 5% overlap with the correct prior distribution (2.176 and 0.692, respectively). The aligned and divergent priors for the intercept mean and slope mean are shown in Figure 6.

It should be noted that the prior specification will interact with misspecification in the model, where it may be that both the prior and the model do not represent the true population parameters. However, this interplay reflects the behavior of an applied researcher, who will likely not alter the prior specification of the intercept and linear slope mean based on the presence or absence of a quadratic slope effect. I estimated the (divergent) population parameter values of the misspecified models to ensure that the divergent priors did not result in posterior estimates that suddenly aligned with these population values. The population values of the intercept and linear slope means were not divergent for the intercept mean (3.1% overlap) and slightly less divergent for the linear slope mean (6.7% overlap).

3.2.8 Data Generation

I chose the number of replications included in each cell of the simulation after assessing at what point the simulation converged. To ensure convergence of the simulation to a stable estimate, I examined cumulative average plots for all conditions. Based on these plots, 1,000 replications were sufficient to ensure that the simulation converged to a stable estimate across all simulation cells. I generated all replications in R (R Core Team, 2019) using the package 'lavaan' (Rosseel, 2012). I simulated data for the following, fully crossed, conditions: population model (3 levels) and sample size (4 levels). These datasets became the starting point for each level of the missing data conditions (i.e., amount of missing data, number of variables affected, location of missing data). I generated missing data as MAR. For the CFA models, each factor's first indicator served as the cause of missingness for the other indicators of that factor that were incomplete. For the LGM, the first time point served as the cause of missingness for the other time points that were incomplete. In addition, if missing values were present in multiple time points, I simulated a dropout pattern (Galbraith et al., 2002; Ortega-Azurduy et al., 2008). In this condition, missing values at a previous time point guaranteed missing values at future time points. I identified new cases with missing

[7] These values are data dependent priors (DDP; Mcneish, 2016), which were selected by first generating 100 samples of $N = 50$ from the population model. These 100 samples were used to estimate the population model with maximum likelihood estimation (MLE). The standard error estimates of the intercept mean and slope mean were averaged across the 100 samples. These values were used to specify the standard deviation hyperparameters of the moderately informative priors. I used the smallest sample size condition to find the standard error estimate with the highest level of uncertainty. The same values will be used across all sample size levels in the main simulation. In applied research, DDPs are controversial, as the researcher technically double-dips by using data to specify the priors that are subsequently used to analyze their data. However, they can aide in model estimation under certain circumstances (e.g., Mcneish, 2016).

values based on their value at time point 1 (i.e., the first time point was the cause of missing data). For the purpose of this dissertation, I solved the following formula to select the incremental increase in the proportion of missing values across all time points:

$$prop = \frac{x \sum_{t=0}^{T} t}{T}. \quad (26)$$

Here, *x* represents the increment with which the proportion of missing values should increase, *T* represents the number of time points with missing values (here, 4), and *prop* represents the desired overall proportion of missing values (here, .15 or .50). This formula can be used to solve for *x* in the following way:

$$x = \frac{prop \times T}{\sum_{t=0}^{T} t}. \quad (27)$$

This meant that, if the overall proportion of missing values was .15, the proportion of missing values increased in increments of .06, resulting in missing value percentages: 0%, 6%, 12%, 18%, 24%. Similarly, if the overall proportion of missing values was .50, the proportion of missing values increased in increments of .20, resulting in missing values percentages: 0%, 20%, 40%, 60%, 80%.

3.2.8.1 Missing Data Generation

I followed a process similar to Enders and Mansolf (2018) to generate missing values in R, using logistic regression to create a missing data indicator for each variable with missing data. I used logistic regression to derive intercept and slope coefficients from the regression of a missing data indicator (*R*) on a standardized predictor:

$$P(R = 1|X) = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}. \quad (28)$$

The intercept value, b_0, determined the amount of missing values predicted by the regression (e.g., an intercept of 0 reflects 50% missing values), while the slope, b_1, reflected the relationship between the predictor and the likelihood of missing data. A positive slope indicates that the probability of missingness increases as the value of the predictor increases. In line with Enders and Mansolf (2018), I selected a slope coefficient that produced a squared correlation of .40, indicating a moderately strong relationship between the cause of missingness and the underlying latent probability for missing data.

After selecting the logistic coefficients, I applied Equation (26) to the actual missing data predictor (i.e., the first indicator of each factor or the first time point), making sure to standardize the variable for the LGM (in the CFA, the first indicator of each item was already on a standardized scale). I used the logit link to obtain a vector of predicted probabilities for each variable that would predict missing values. I then used a binomial distribution function to generate a vector of missing data indicators for each variable that would contain missing values, where the predicted probabilities from the previous steps defined the distribution's probability of success. This process continued until the desired amount of missingness was reached. I coded each of the values within a variable as missing if its corresponding indicator equaled one.

3.2.9 Bayesian Estimation

Bayesian estimation was done through the program M*plus* (Muthén & Muthén, 1998-2017). The CFA models and the LGMs with diffuse priors were all estimated using the M*plus* default priors (see footnote 7, p. 36). All analyses generated four MCMC chains using M*plus*' implementation of the Gibbs sampler. Each chain consisted of 20,000 iterations, with the first 10,000 discarded as burn-in. This number of iterations was selected after testing several chain lengths for a select number of replications in each cell and inspecting the trace plots, the $\hat{R}$ convergence diagnostic, and ensuring that the each parameter's effective sample size (ESS) was > 1000 (Zitzmann & Hecht, 2019). To further ensure that convergence was obtained across all replications, the $\hat{R}$ convergence diagnostic and effective sample size were checked for all replications across all conditions.

3.2.10 Outcomes of Interest

I assessed the performance of the model fit indices in two main ways. First, for the PPP-value, BRMSEA, BCFI, and BTLI, the model fit index values were compared across misspecification levels to assess whether they systematically worsened as misspecification became more severe. For these indices, it was also possible to assess whether the average fit index value indicated good model fit using cutoff values: PPP > .05, BRMSEA < .06, BCFI and BTLI > .95. Furthermore, M*plus* reports 90% credible intervals for the posterior distribution of the BRMSEA, BCFI, and BTLI.[8] As discussed in the Introduction, these intervals can provide additional insight into a model's approximate fit. I used the cutoff values listed above in the current dissertation for practical reasons: Those cutoff values are generally used in applied research and were also referenced in the study introducing the implementation of these indices to M*plus* (Asparouhov & Muthén, 2020). Although the cutoff value for the PPP of .05 may mirror the frequentist *p*-value significance cutoff (for an alpha level of .05), it should not be interpreted in the same manner. As was discussed in the Introduction, a model that fits the data well is expected to have a PPP-value close to 0.5 (i.e., half of the replicated datasets had greater discrepancy values compared to the observed data). A misspecified model is expected to have a PPP-value close to 0 (i.e., most of the replicated datasets had greater discrepancy values compared to the observed data). The cutoff value of .05 serves the same purpose as cutoff values for the approximate fit indices: to ease interpretation.

Second, I will assess how often the model fit and selection indices select the correctly specified model over the misspecified models. This will provide insight into which model fit and selection indices can be used for model selection.

---

[8] During the pre-proposal meeting, we discussed looking at additional intervals, such as a 90% HPD interval. However, M*plus* does not provide the full posterior distribution of the approximate fit indices and only reports the 90% credible interval. Thus, this extension of the results could not be implemented in Study1.

3.3 Results

The results are organized as follows: I first discuss findings regarding convergence of each of the replications in the simulation. Next, I examine the results for each population model sequentially.

3.3.1 Convergence

For each replication in the simulation, I extracted largest $\hat{R}$ and smallest ESS to assess convergence and precision of the posterior distributions. A maximum $\hat{R} < 1.05$ and minimum ESS > 1000 indicated that a replication converged. According to this criterium, all replications of the CFA-Simple and LGM population models were converged. For the CFA-Complex model, I found that, for a small minority of replications, the estimation resulted in a non-positive definite **Φ** matrix. This issue occurred more often when the sample size was small. For $n = 50$, 242 out of the total 30,000 replications (0.8%) did not converge, for $n = 100$, just 48 out of the total 30,000 replications (0.16%) did not converge, and for $n = 250$, a mere 3 out of the total 30,000 replications (0.01%) did not converge. All replication converged for the $n = 500$ sample size. Non-convergence occurred only for misspecified models and was more likely if 50% of values were missing for nine items. Non-convergence did not appear related to the location of the misspecification. Only replications for which all model specifications converged were included in the results presented in Section 3.3.3. In applied research, this issue could be resolved by specifying a weakly informative prior on the latent factor covariance matrix, such as $IW(\mathbf{I}, p+1)$, where p stands for the number of latent factors.[9]

3.3.2 CFA-Simple

3.3.2.1 The Value of the Model Fit Indices

Model fit indices should increasingly indicate that a model fits the data poorly as model misspecification becomes more severe. I used boxplots to assess whether the model fit indices followed this pattern for the CFA-Simple population model. Within each figure, the rows represent the sample size levels, and the columns represent the missing data conditions. Specifically, the complete data conditions are shown in the left-most column, while the remaining four columns focus on the conditions that are increasingly affected by missing values. Within each plot, the different model specifications are compared through boxplots based on the observed fit index values across all replications.

Results for the PPP-value appear in Figure 7. Each plot includes a horizontal line at PPP = .05 to emphasize at what point a model might be rejected based on this cutoff value. Across all conditions, the PPP-value decreased as model misspecification became more severe. However, sample size greatly affected to what extend the PPP-value decreased. Specifically, for $n = 50$ and 100, the PPP-value was unlikely to drop below .05, even for the most severe level of model misspecification. In contrast, for the largest sample size ($n = 500$), the PPP-value was sensitive to moderate model misspecification. Missing values reduced the ability of the PPP-value to differentiate

[9] Reference: http://www.statmodel.com/discussion/messages/11/1602.html?1597708138.

between different levels of model misspecification. For example, the PPP-value decreased to a similar extent for complete data with a sample size of 100 as it did for data with 50% missing values in nine items with a sample size of 250.

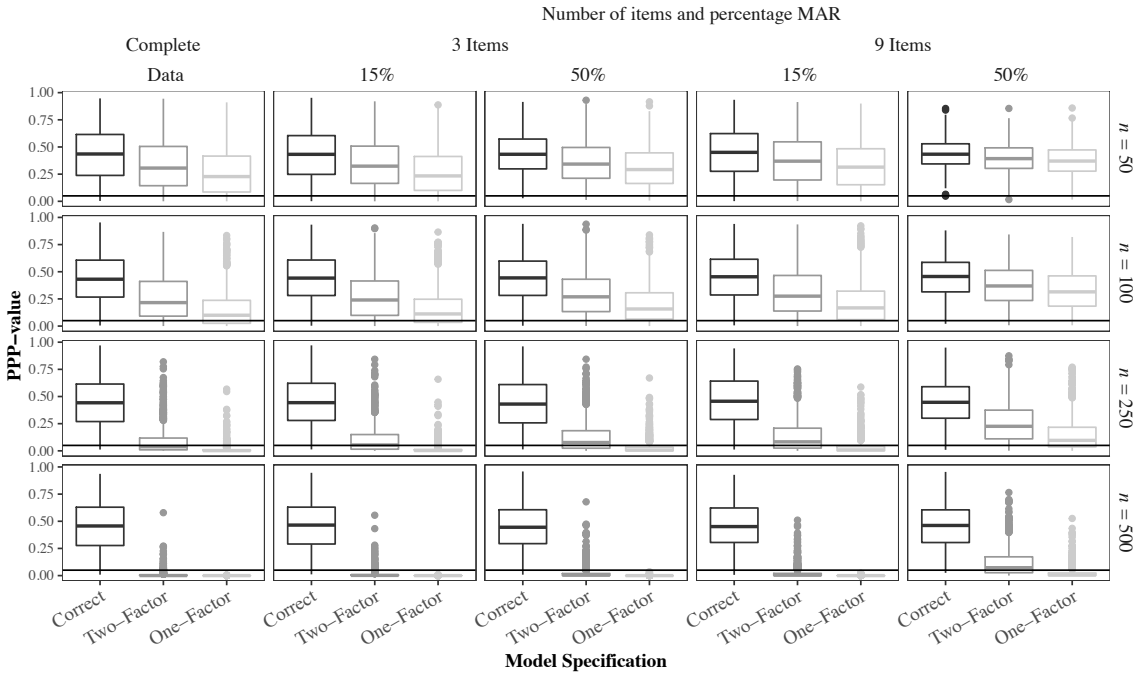


Figure 7. CFA-Simple: PPP-value across simulation conditions.

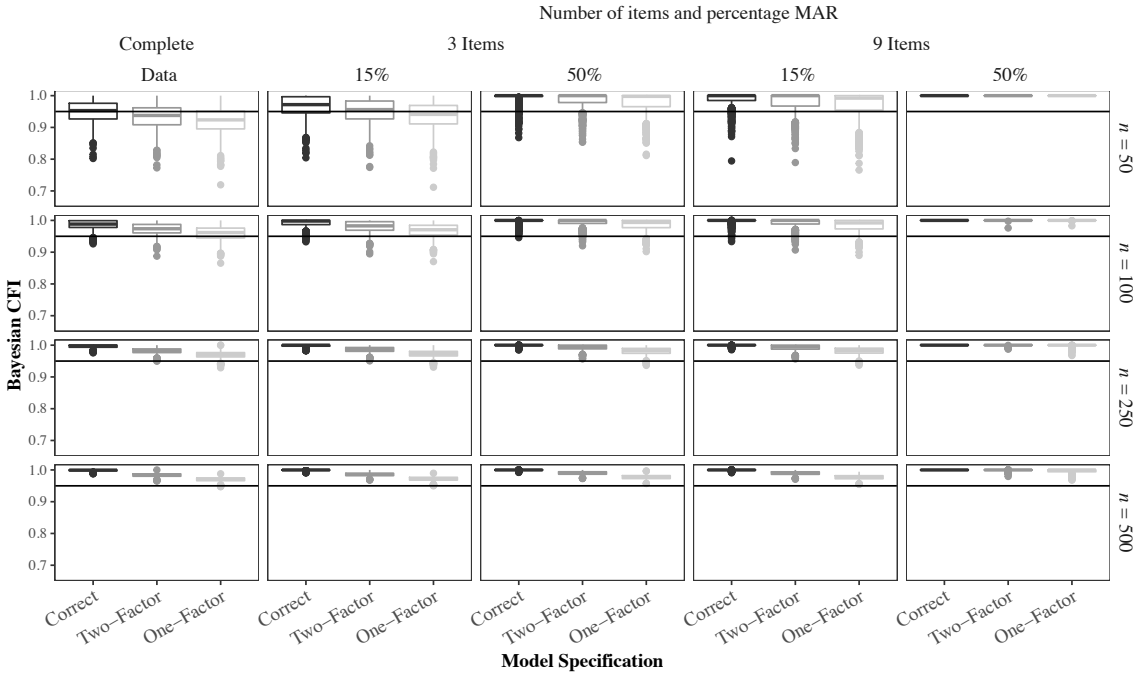


Figure 8. CFA-Simple: BCFI across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

Results for the BCFI and BTLI appear in Figures 8 and 9. Each plot includes a horizontal line at .95 to emphasize at what point a model might be rejected based on this cutoff value. As they follow a similar pattern, only the BCFI (Figure 8) will be discussed here. In this figure, we see that, although the BCFI decreased as model misspecification became more severe, its value was unlikely to fall below .95. The BCFI increased as the sample size increased (moving down the plots within one column). Unexpectedly, the BCFI also increased and varied less across replications as the number of missing values increased (moving left to right within a row).

Most notably, the BCFI was equal to 1 across all replications and model specifications when n = 50 with 50% missing values in nine items (top-right plot). Closer inspection of these analyses showed that the models were converged, and their associated PPP-value was interpretable. It appears that, when the sample size is this small (and further reduced within variables due to missing values), D_s^{obs} is smaller than p^* resulting in a BCFI estimate that is greater than 1. These estimates are adjusted to 1 in the output generated by M*plus*. As discussed in Section 3.2.1.2, p^* is related to the number of observed variables. For the CFA-Simple model, p^* is 135. It may be that this model includes too many variables relative to the information provided through the sample observations and that this is preventing accurate estimation of the BCFI (and similarly the BTLI and BRMSEA). This issue was also briefly mentioned by Asparouhov and Muthén (2020), who explained that the difference between the baseline model discrepancy, $D_{B,s}^{obs}$, and the estimated model discrepancy, D_s^{obs}, becomes too small to evaluate approximate model fit.

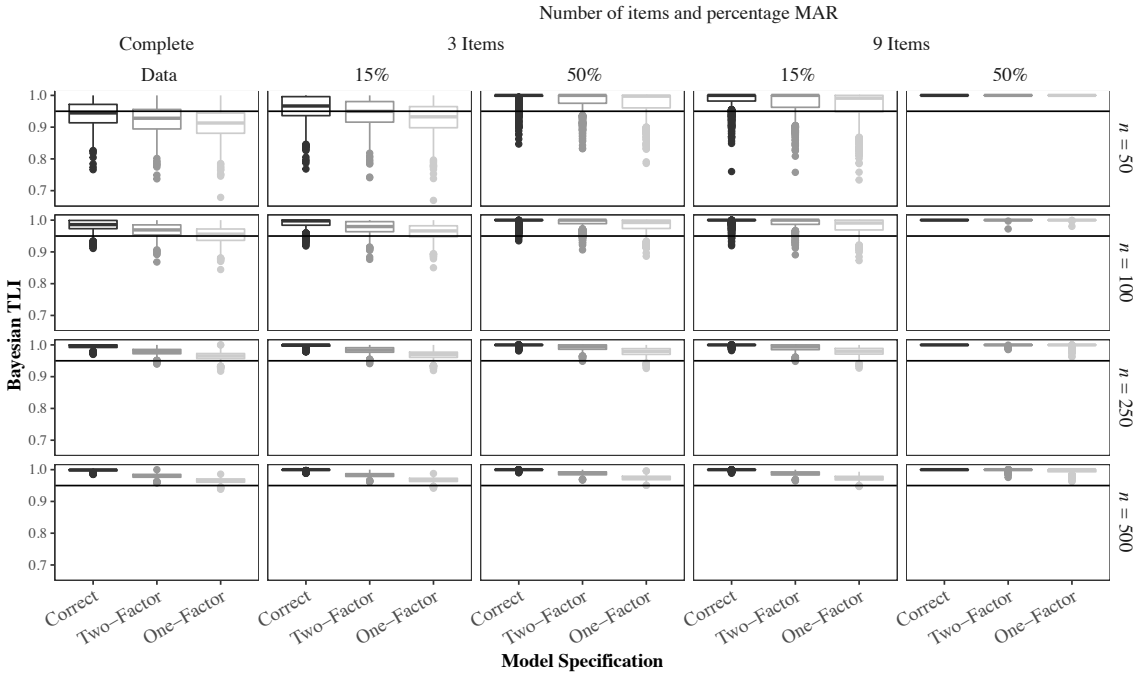


Figure 9. CFA-Simple: BTLI across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

Figure 10 depicts results for the BRMSEA. Each plot includes a horizontal line at .06 to emphasize at what point a model might be rejected based on this cutoff value. The pattern of results is similar to the BCFI and BTLI; although the BRMSEA did increase (reflecting a decline in model fit) as the model misspecification became more severe, it tended to move towards lower values as the sample size increased. For n = 500, the BRMSEA was unlikely to indicate poor model fit, even for the most severe model misspecification level. Mirroring the BCFI and BTLI, the BRMSEA also decreased and varied less across replications as the number of missing values increased.

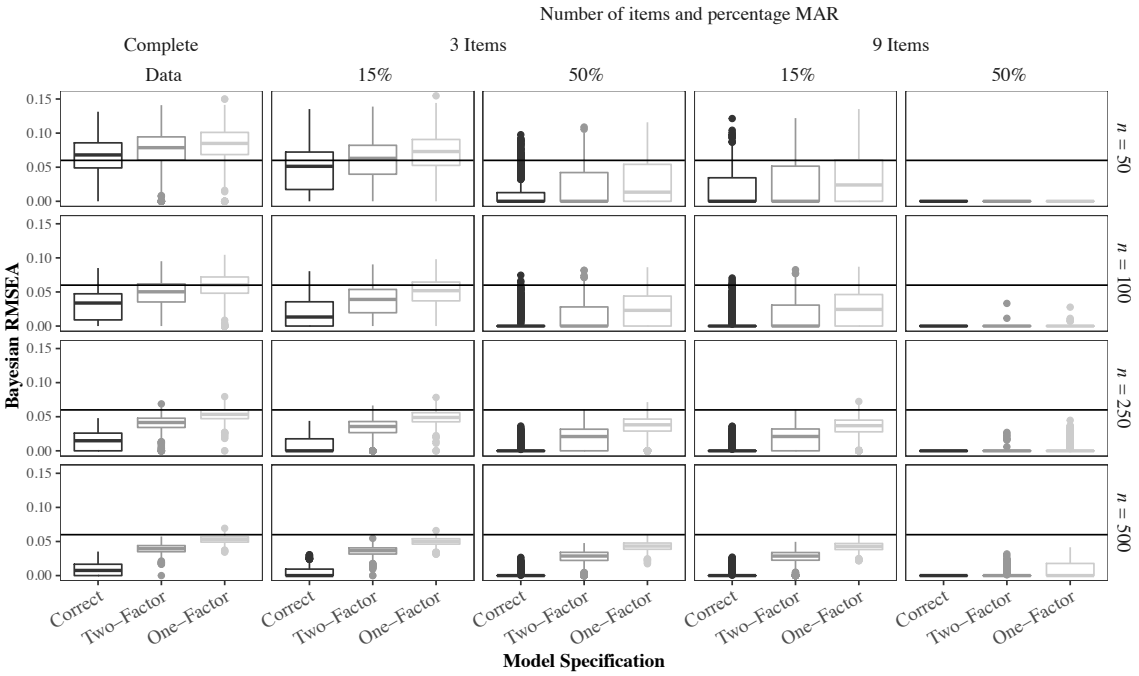


Figure 10. CFA-Simple: BRMSEA across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

3.3.2.2 Does the Model Fit the Data well?

An important question for researchers is whether a particular model fits the data well. The PPP-value, BCFI, BTLI, and BRMSEA can all be used to answer this question by using cutoff values or (for the approximate fit indices) using a credible interval. I will first go over the results based on cutoff values, after which I will present the result based on credible intervals.

3.3.2.2.1 Using Cutoff Values

Figure 11 presents the proportion of replications that are rejected based on a given fit index's cutoff value. This figure follows the same layout as Figures 7-10. However, instead of boxplots, the plot includes lines showing the proportion of times each fit index rejected a particular model specification. Ideally, each line starts at 0 for the correctly specified model, after which it should steeply increase to 1 for the misspecified models.

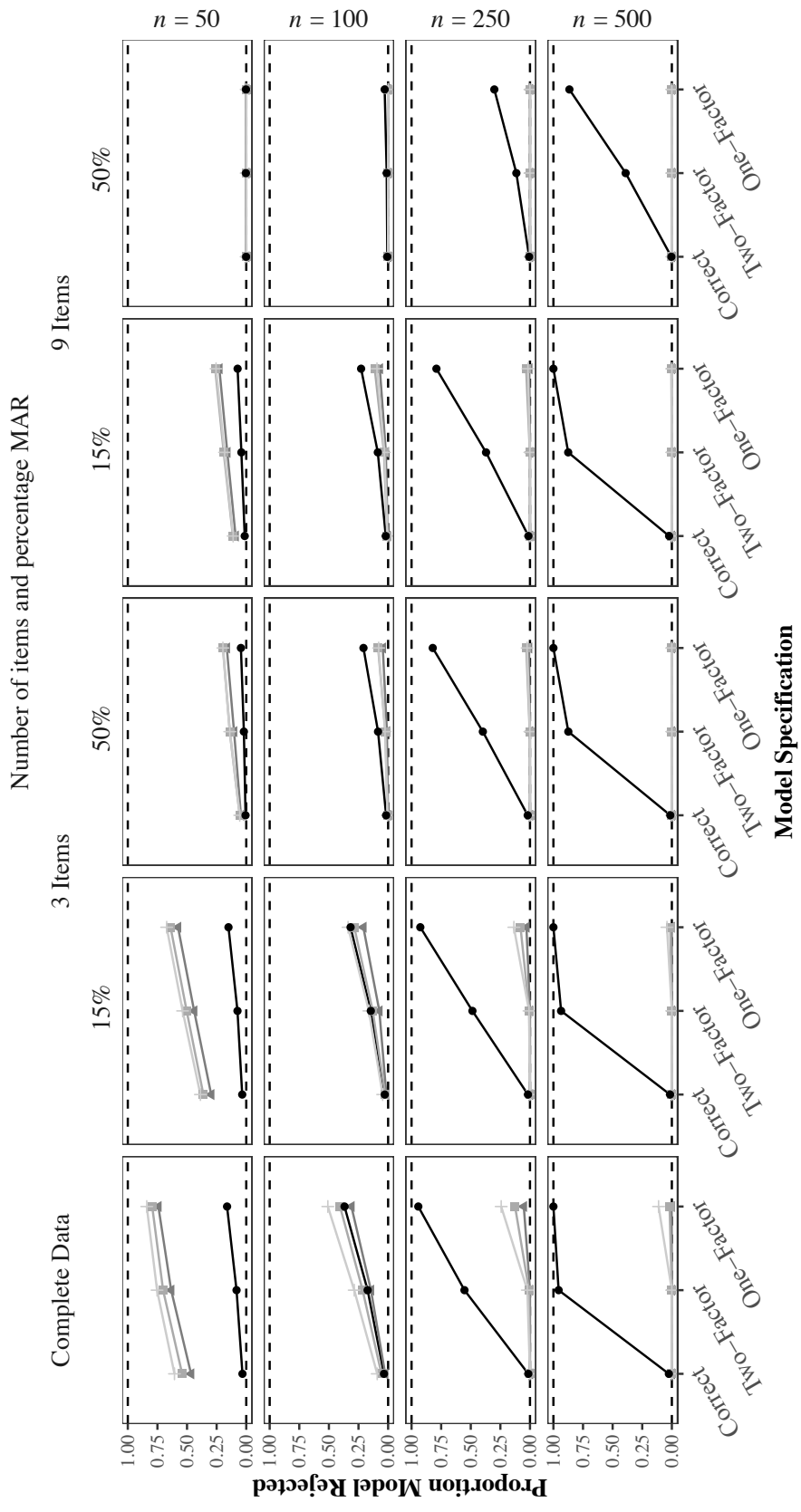


Figure 11. CFA-Simple: Proportion of times a model was rejected based on each fit index's cutoff value across simulation conditions.

In general, the fit indices rejected misspecified models more often if data were complete compared to data containing missing values. Overall, the PPP-value was most sensitive to model misspecification. However, the sample size affected the extent to which misspecified models were rejected. For n = 500 (bottom row), the PPP-value was highly sensitive and likely to reject a moderately misspecified model, particularly if data were complete. In contrast, for n = 50, the PPP-value was still very likely to accept a severely misspecified model. The sensitivity of the PPP-value and its cutoff value to misspecification was also lower if more variables had missing values and if a higher percentage of values was missing. The PPP-value's sensitivity to the sample size is not unexpected, given that the discrepancy function at the foundation of the PPP-value is rooted in the Chi-square statistic.

Results for the BCFI, BTLI, and BRMSEA followed a similar pattern. Overall, the approximate fit indices were more likely to reject a severely misspecified model than a correctly specified model. However, even under conditions in which the approximate fit indices had the highest rejection rates for misspecified models, those rates were far below the ideal of 1. The tendency of the approximate fit indices to retain a misspecified model increased as the sample size increased. In contrast, the approximate fit indices became more likely to reject a correctly specified model for samples of n = 50 (top row), particularly if data were complete. As the number of variables with missing data and percentage of missing values increased, any differences between sample sizes disappeared and all indices were likely to retain the model across all levels of misspecification.

3.3.2.2.2 Using 90% Credible Intervals

For the approximate fit indices, it is possible to use their 90% credible interval to assess model fit. This approach results in one of three conclusions: (1) the model fits the data well (the entire 90% credible interval is beyond the cutoff that denotes good fit), (2) model fit is inconclusive (the cutoff values is within the 90% credible interval), (3) the model does not fit the data (the entire 90% credible interval is < .95 for the BCFI/BTLI or > .06 for the BRSMEA). The results for the BCFI, BTLI, and BRMSEA based on this approach appear in Figures 12, 13, and 14, and are organized in the same manner as Figures 7-10. However, in the figures below, the stacked bars represent the proportion of replications that resulted in one of the three conclusions (good fit; inconclusive fit; poor fit) for each model specification. Ideally, the entire bar is the lightest shade for the correctly specified model and the darkest shade for the misspecified models. As the pattern of results is similar for all three indices, only the results for the BCFI (Figure 12) are discussed in detail.

Overall, the conclusions drawn based on the credible interval followed those based on the cutoff value: as the sample size or amount of missing values increased, the 90% credible intervals increasingly indicated that the model fit the data well, regardless of the level of model misspecification. However, the figures also illustrate some scenarios for which the 90% credible interval method would result in a more nuanced conclusion regarding model fit. Specifically, for small sample sizes with minimal amounts of missing values, the 90% credible intervals often indicated that model fit was inconclusive. This classification is important diagnostic information that could help a

researcher realize that they may not be able to rely on the point estimates of the approximate fit indices for assessing model fit. However, as this effect disappeared with larger sample sizes, it may be difficult to know when this reasoning can be applied.

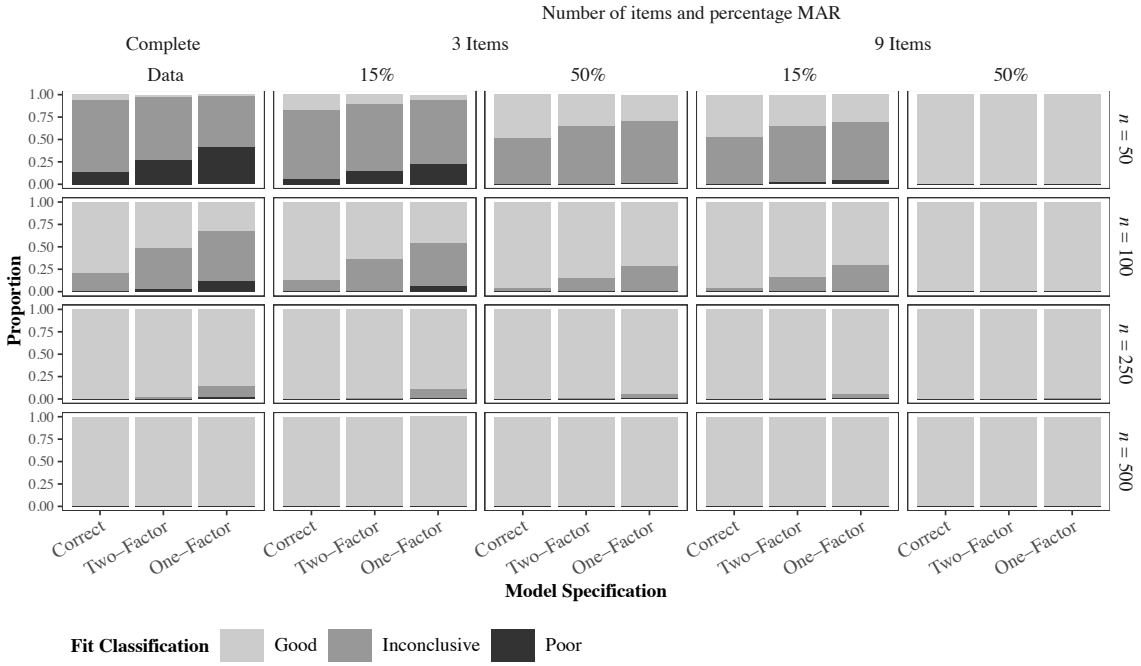


Figure 12. CFA-Simple: Model fit classification based on 90% BCFI credible interval.

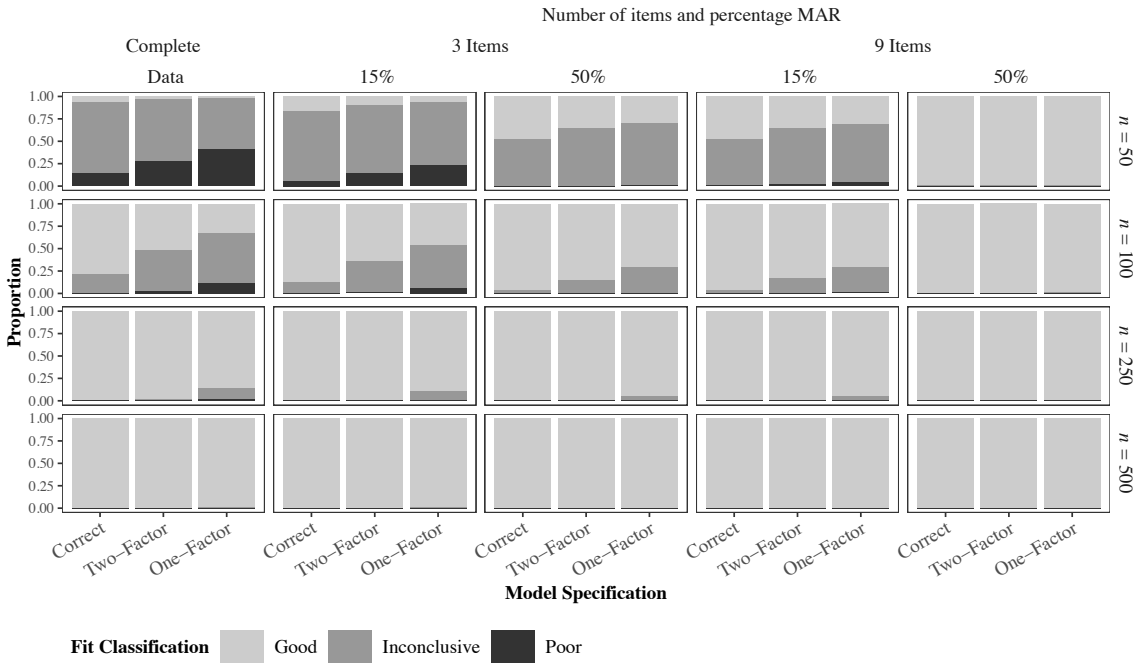


Figure 13. CFA-Simple: Model fit classification based on 90% BTLI credible interval.

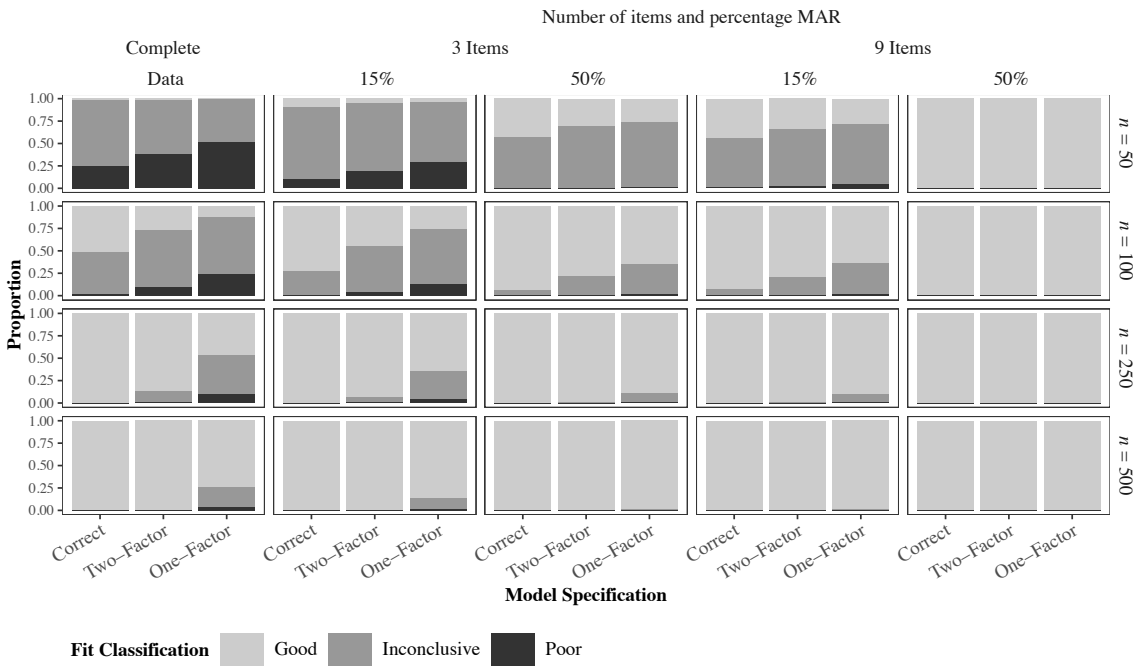


Figure 14. CFA-Simple: Model fit classification based on 90% BRMSEA credible interval.

3.3.2.3 Model Selection

Model fit and selection indices can also be used for model selection. Figure 15 presents the proportions of replications selected based on a given fit index's value. This figure follows the same layout as Figure 11. Ideally, each line starts at 1 for the correctly specified model, after which it should steeply decrease to 0 for the misspecified models. Model selection is only possible if the model fit or selection index changes across model specifications. Recall that the results reported in Section 3.3.2.1 showed that the approximate fit indices tended to be equal to 1 across all replications and model specifications with small sample sizes or large numbers of missing values. Thus, before examining model selection across fit indices, I first examined to what extent the approximate model fit indices were equivalent (and equal to 1 for BCFI/BTLI or 0 for BRMSEA) across model specifications. The proportions of replications that resulted in equivalent fit index values across model specifications are reported in Table 2. From this Table, it becomes clear that the issue of equivalent approximate model fit indices was present across all sample sizes and missing value conditions. However, the approximate model fit indices were much more likely to be equivalent if 50% of values are missing in 9 items, even if the overall sample size was 500. As the sample size increased, the fit indices became less likely to be equivalent with complete data or with moderate numbers of missing values.

The model selection proportions for the approximate fit indices reported in Figure 15 are based solely on replications for which the fit indices were different across model specifications. This means that for the right-most column in the figure, displaying the

conditions with the largest number of missing values, the points and lines are based on just a few replications. Thus, those results should be interpreted with caution.

Table 2. CFA-Simple: Proportion of replications for which the approximate model fit indices were equal across all three model specifications.

Sample Size	Complete Data	Number of items and percentage MAR			
		3 items		9 items	
		15%	50%	15%	50%
50	0.033	0.132	0.581	0.542	1.000
100	0.072	0.171	0.620	0.577	0.998
250	0.016	0.039	0.290 *	0.290 *	0.991
500	0.001	0.002	0.037	0.035	0.916

* For the BRMSEA these proportions were equal to 0.292 instead of 0.290.

Overall, all fit indices were likely to select the correctly specified model for $n = 250$ and 500. For $n = 250$, the PPP-value (black dot) and the BIC (grey box with a cross) were slightly more sensitive to the presence of missing values. Specifically, with 50% missing values in 9 items, the PPP-value and DIC were slightly less likely to select the correctly specified model, and slightly more likely to select the moderately misspecified model. For $n = 100$ and $n = 50$, the performance of the fit indices diverged more clearly. Overall, the BIC was least likely to select the correctly specified model, particularly for $n = 50$ or when data contained missing values. Specifically, for $n = 50$ with 50% missing values in nine items, the BIC was more likely to select either of the misspecified models than the correctly specified model. The PPP-value was less sensitive to missing data than the BIC but was less likely to select the correct model than the DIC if missing data was limited. Perhaps surprisingly, the approximate fit indices were most likely to select the correctly specified model across conditions when $n = 50$ or 100. It should be noted that these indices are not included in the plot for $n = 50$ with 50% missing values in 9 items (top-right plot) as their values were equal to 1 (or 0 for the BRMSEA) across all three model specifications (Table 2). Thus, as long as the approximate fit indices are properly estimated, they will most likely select the correctly specified model across all included conditions.

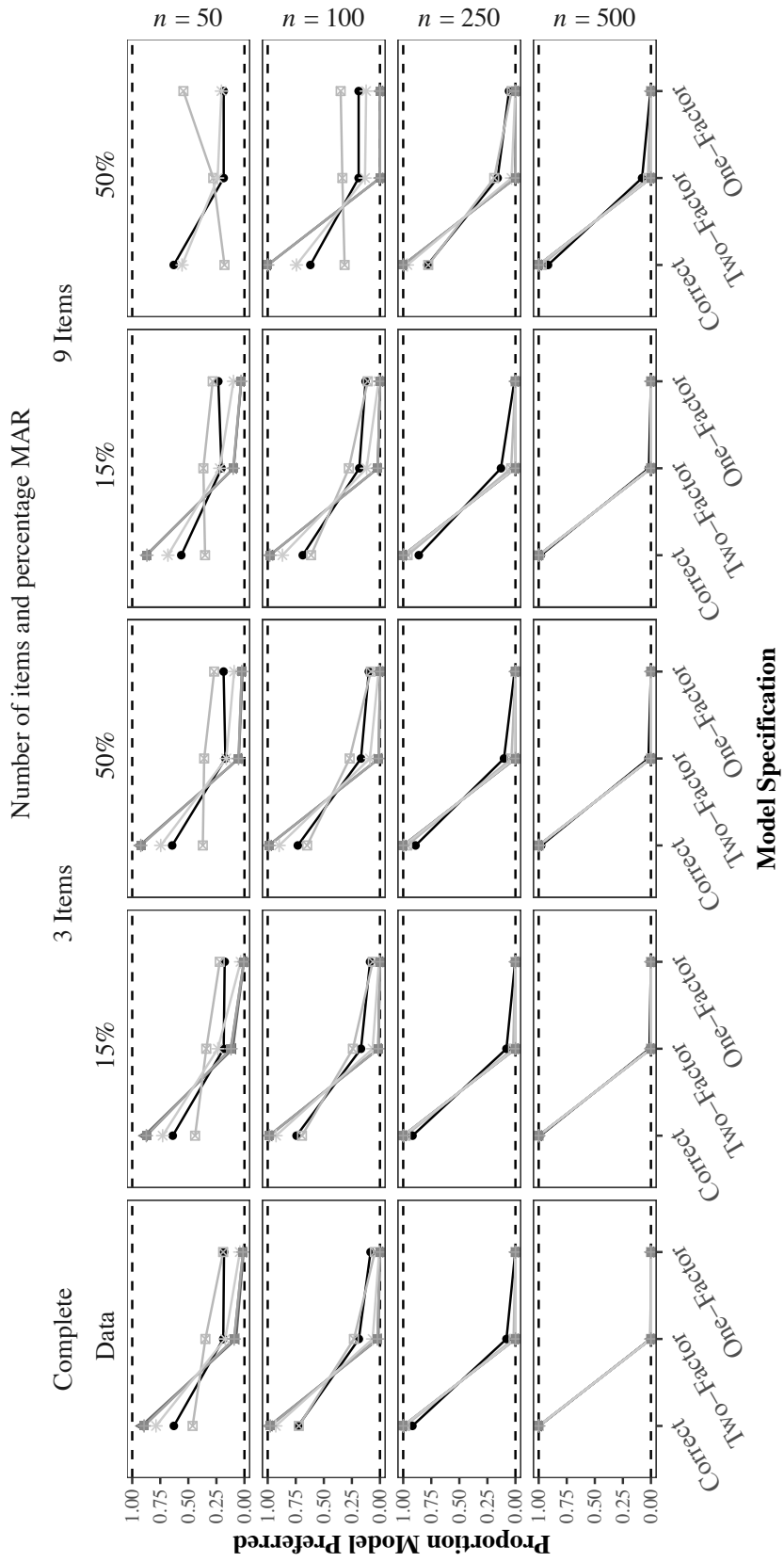


Figure 15. CFA-Simple: Proportion of times a model was selected based on each fit index's value across simulation conditions.

3.3.3 CFA-Complex

The location of the missing values relative to the location of the misspecification did not affect the results for the CFA-Complex population model. Thus, results presented here are based on the simulation conditions in which the variables with missing values were not involved in the model misspecification. Results of the condition in which variables with missing values were involved in the misspecification can be found on the OSF page.

3.3.3.1 The Value of the Model Fit Indices

Model fit indices should increasingly indicate that a model fits the data poorly as model misspecification becomes more severe. As with the CFA-Simple population model, I used boxplots to assess whether the model fit indices followed this pattern for the CFA-Complex population model.

Results for the PPP-value appear in Figure 16. Each plot includes a horizontal line at PPP = .05 to emphasize at what point a model might be rejected based on this cutoff value. Across all conditions, the PPP-value decreased as model misspecification became more severe. However, sample size greatly affected to what extend the PPP-value decreased. Specifically, for n = 50, the PPP-value was unlikely to drop below .05, even for the most severe level of model misspecification. In contrast, for the largest sample size (n = 500), the PPP-value dropped below .05 as soon as one cross-loading was omitted from the specification. Missing values reduced the ability of the PPP-value to differentiate between different levels of model misspecification. For example, the PPP-value decreased to a similar extent for complete data with a sample size of 50 as it did for data with 50% missing values in nine items with a sample size of 100.

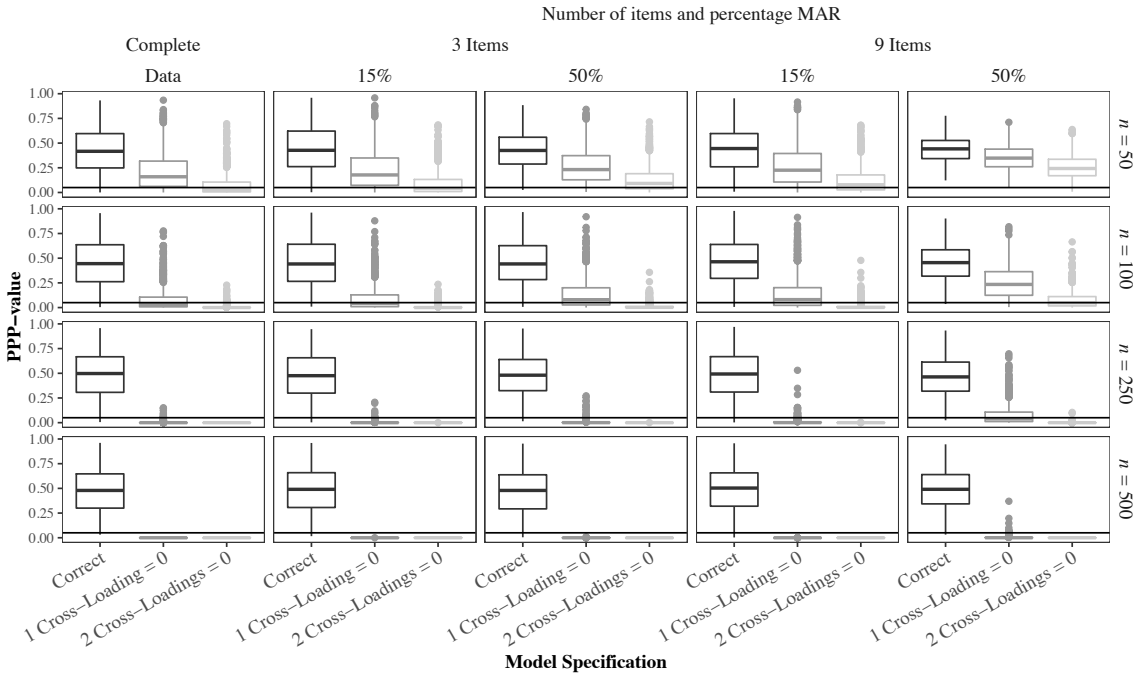


Figure 16. CFA-Complex: PPP-value across simulation conditions.

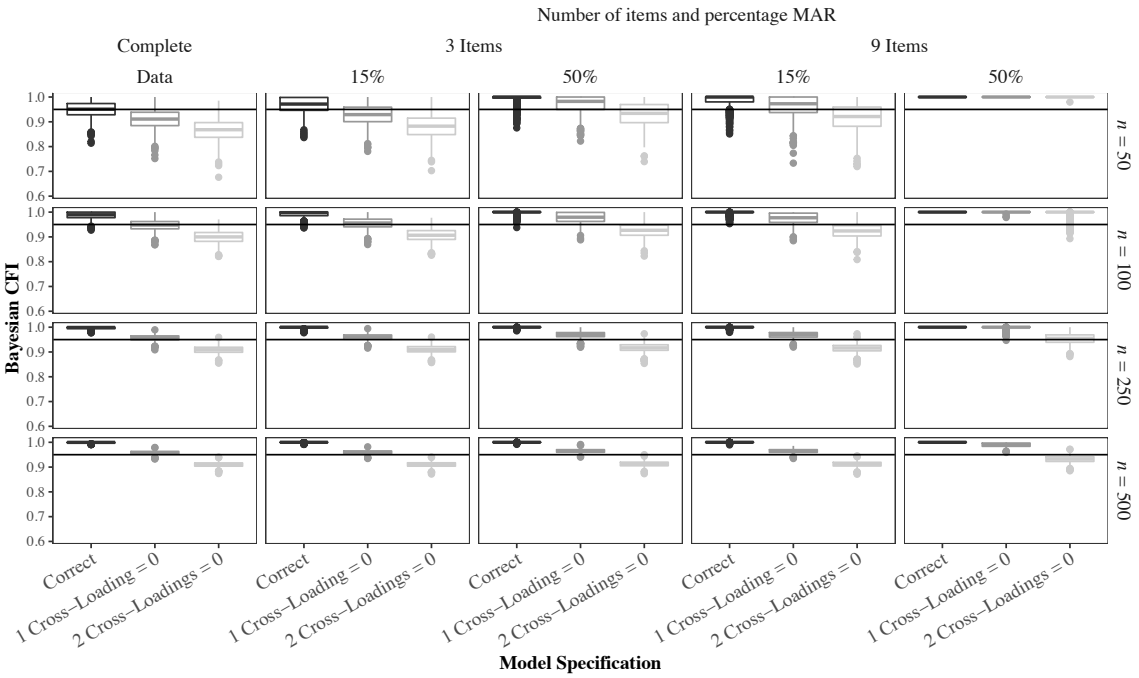


Figure 17. CFA-Complex: BCFI across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

Results for the BCFI and BTLI appear in Figures 17 and 18. Each plot includes a horizontal line at .95 to emphasize at what point a model might be rejected based on this cutoff value. As they follow a similar pattern, I discuss only the BCFI (Figure 17) here. In this figure, we see that, although the BCFI decreased as model misspecification became more severe, its value was not as sensitive to model misspecification as the PPP-value. In contrast to the pattern of results for the CFA-Simple model, the BCFI did not appear to increase as the sample size increased (moving down the plots within one column). Similar to the pattern of results for the CFA-Simple model, the BCFI increased and varied less across replications as the number of missing values increased (moving left to right within a row). The BCFI was equal to 1 across all replications and sample sizes when the model was correctly specified with 50% missing values in nine items (top-right plot). As with the CFA-Simple model, closer inspection of these analyses showed that the models were converged, and their associated PPP-value was interpretable. The same explanation as found for the CFA-Simple model appears to be applicable here. It may be that this model includes too many variables relative to the information provided through the sample observations and that this is preventing accurate estimation of the BCFI (and similarly the BTLI and BRMSEA). What is surprising is that this behavior also occurred for the larger sample sizes included in the simulation and that it only occurred for the correctly specified model (i.e., the model with the largest number of freely estimated parameters).

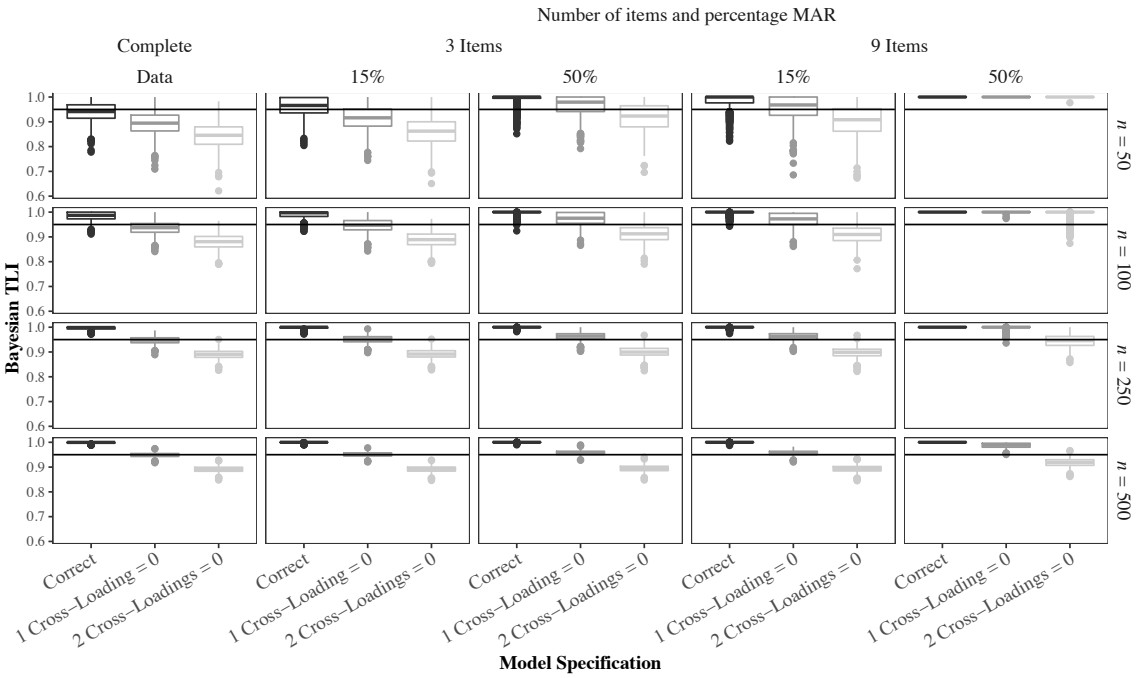


Figure 18. CFA-Complex: BTLI across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

Results for the BRMSEA appear in Figure 19. Each plot includes a horizontal line at .06 to emphasize at what point a model might be rejected based on this cutoff value. The pattern of results is similar to the BCFI and BTLI; although the BRMSEA did increase (reflecting a decline in model fit) as the model misspecification became more severe, it tended to move towards lower values as the number of missing values increased (moving left to right within a row). For n = 50, if 50% of values were missing in 9 items, the BRMSEA appeared to reflect perfect fit across levels of misspecification. This pattern of results barely improved as the overall sample size increased.

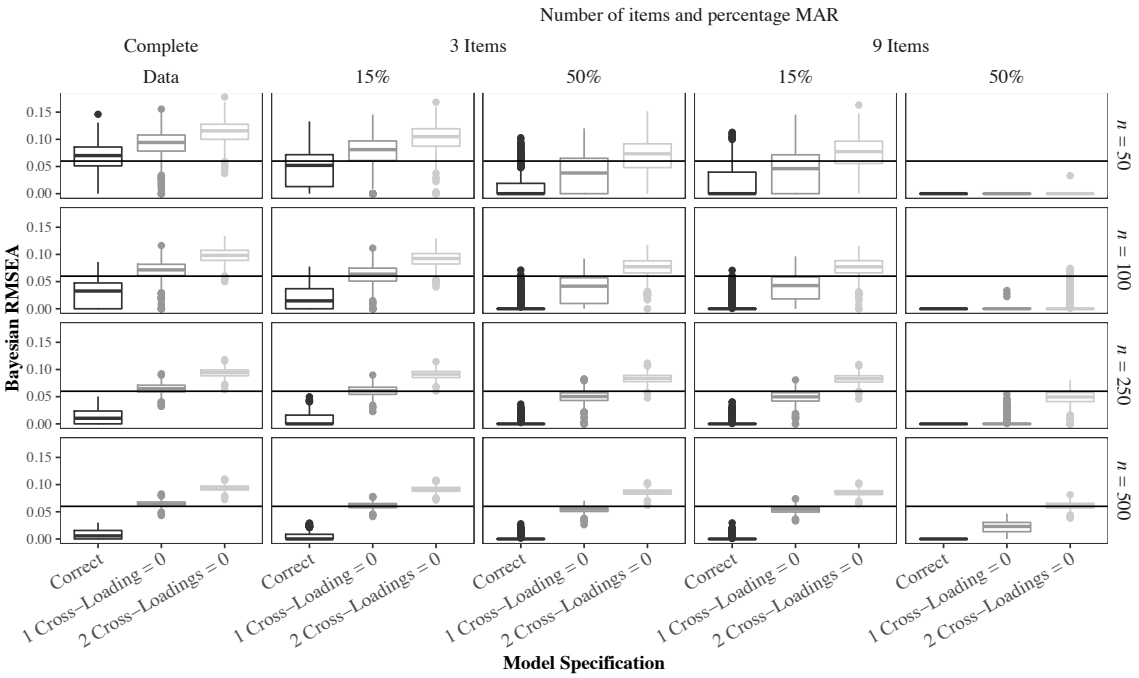


Figure 19. CFA-Complex: BRMSEA across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

3.3.3.2 Does the Model fit the Data well?

An important question for researchers is whether a particular model fits the data well. The PPP-value, BCFI, BTLI, and BRMSEA can all be used to answer this question by using cutoff values or (for the approximate fit indices) using a credible interval. I will first go over the results for the CFA-Complex based on cutoff values, after which I will present the result based on credible intervals.

3.3.3.2.1 Using Cutoff Values

Figure 20 presents the proportion of replications that are rejected based on a given fit index's cutoff value. This figure follows the same layout as Figure 11. Ideally, each line starts at 0 for the correctly specified model, after which it should steeply increase to 1 for the misspecified models.

In general, the fit indices rejected misspecified models more often if data were complete compared to data containing missing values. For $n = 50$ with 50% missing values in 9 items (top-right panel), none of the indices rejected any of the model specifications. Overall, the PPP-value was most sensitive to model misspecification. However, the sample size affected the extent to which misspecified models were rejected. For $n = 500$ (bottom row), the PPP-value was highly sensitive and likely to reject a moderately misspecified model, particularly if data were complete. In contrast, for $n = 50$, the PPP-value was still very likely to accept a severely misspecified model. The sensitivity of the PPP-value and its cutoff value to misspecification was also lower if more variables had missing values and a higher percentage of values was missing.

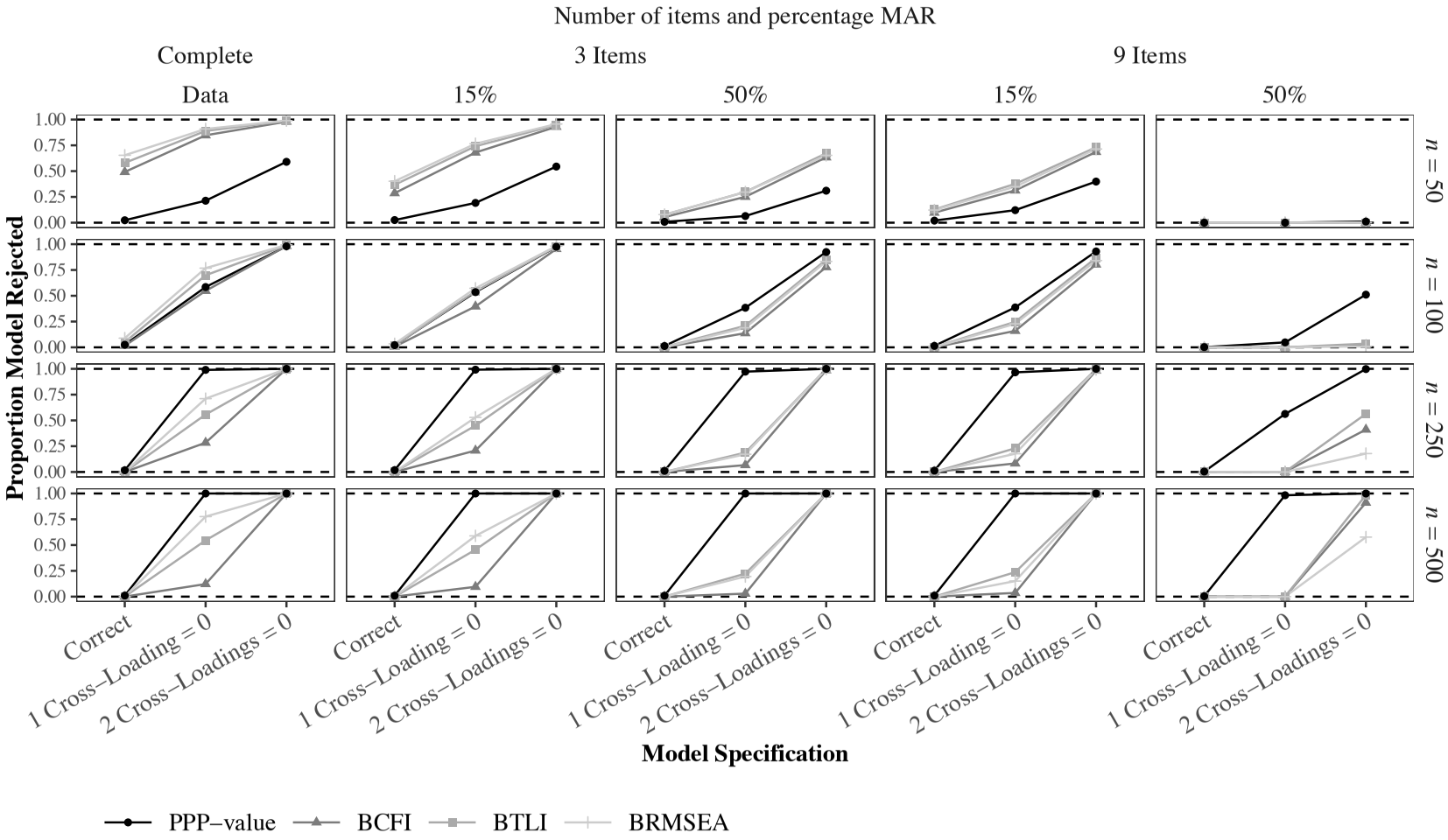


Figure 20. CFA-Complex: Proportion of times a model was rejected based on each fit index's cutoff value across simulation conditions.

Results for the BCFI, BTLI, and BRMSEA followed a similar pattern if n = 50 or 100. For n = 50, the approximate fit indices were likely to reject the correctly specified model if data were complete. As the presence of missing data increased, the approximate fit indices became less likely to reject a model, even if both cross-loadings were omitted from the specification. For n = 100, the approximate fit indices were more likely to reject a model that omitted both cross-loadings, although the proportion of rejected models still declined as the number of missing values increased.

For n = 250 and 500, the approximate fit indices all rejected the most severely misspecified model, regardless of the presence of missing data. However, if the estimated model omitted only one cross-loading, their likelihood to reject the model diverged, particularly when data were complete (left-most column). It appears that the BRMSEA and the BTLI (with their cutoff values) are somewhat more sensitive to model misspecification than the BCFI. As the number of missing values increased, none of the approximate fit indices became likely to reject the moderately misspecified model.

3.3.3.2.2 Using 90% Credible Intervals

For the approximate fit indices, it is possible to use their 90% credible interval to assess model fit. This approach results in one of three conclusions: (1) the model fits the data well (the entire 90% credible interval is beyond the cutoff that denotes good fit), (2) model fit is inconclusive (the cutoff values is within the 90% credible interval), (3) the model does not fit the data (the entire 90% credible interval is < .95 for the BCFI/BTLI or > .06 for the BRSMEA). The results for the BCFI, BTLI, and BRMSEA based on this approach appear in Figures 21, 22, and 23, organized in the same manner as Figures 12-14. Ideally, the entire bar is the lightest shade for the correctly specified model and the darkest shade for the misspecified models. As the pattern of results is similar for all three indices, only the results for the BCFI (Figure 21) are discussed in detail.

Overall, the credible intervals were more likely to indicate poor model fit for the most severe model misspecification (omitting both cross-loadings) as the sample size increased. In contrast, for the moderate level of misspecification, the credible intervals became more likely to indicate good model fit as the sample size increased. In both instances, the credible intervals were more likely to indicate inconclusive model fit as the sample size decreased. The presence of missing values increased the probability that the credible intervals would indicate good or inconclusive model fit. If 50% of values were missing in 9 items, the credible intervals were unlikely to indicate poor model fit (except for n = 500). This pattern can be explained by the difficulty in computing the approximate model fit indices (and thus their credible intervals) due to the large number of estimated parameters relative to the observed sample size.

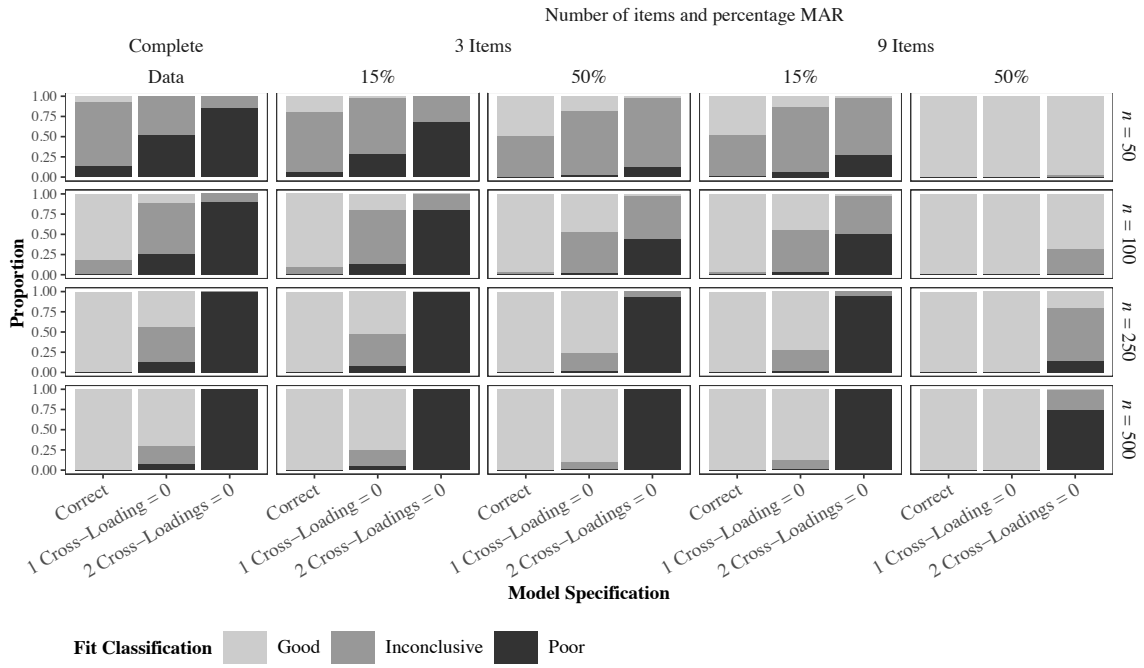


Figure 21. CFA-Complex: Model fit classification based on 90% BCFI credible interval.

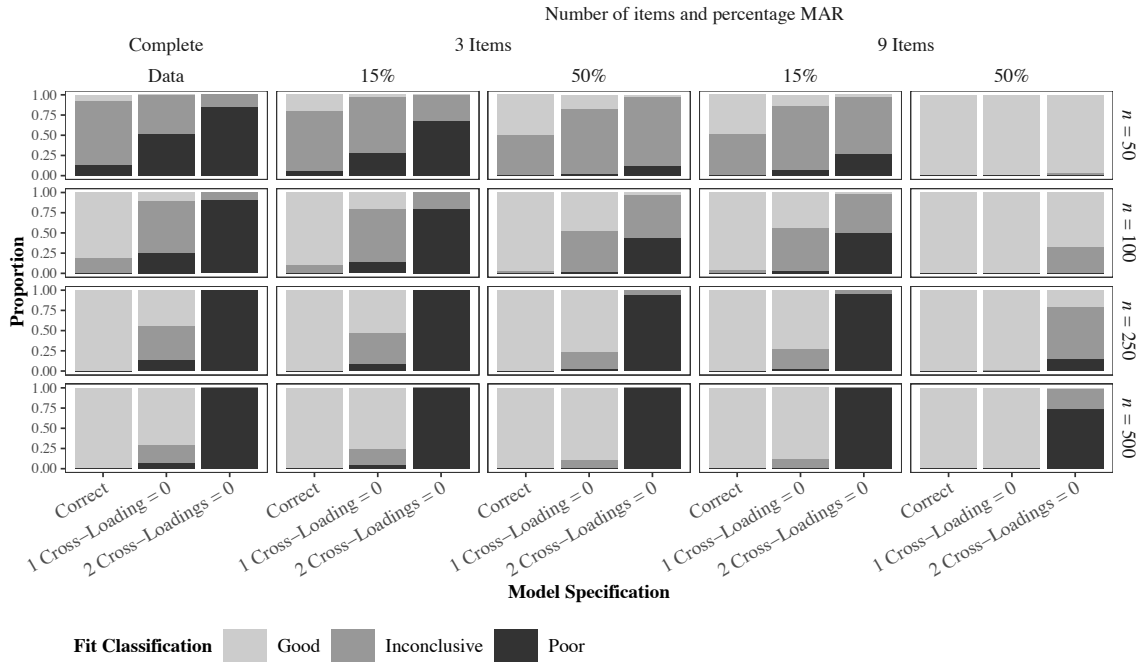


Figure 22. CFA-Complex: Model fit classification based on 90% BTLI credible interval.

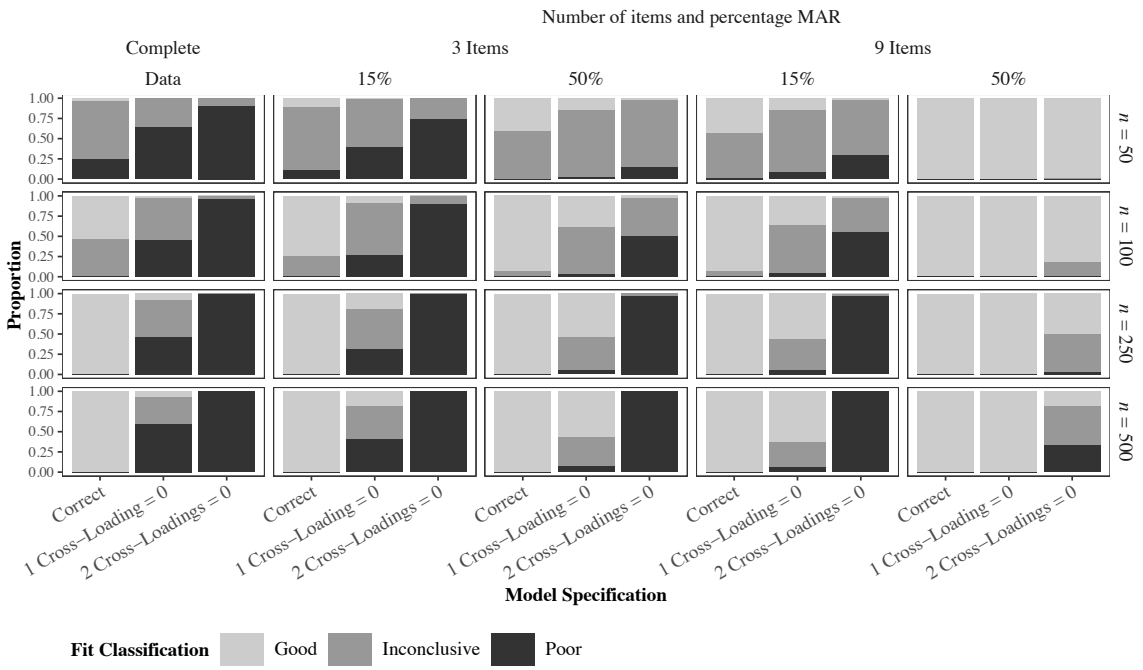


Figure 23. CFA-Complex: Model fit classification based on 90% BRMSEA credible interval.

3.3.3.3 Model Selection

Model fit and selection indices can also be used for model selection. Figure 24 shows the proportions of replications selected based on a given fit index's value. This figure follows the same layout as Figure 15. Ideally, each line starts at 1 for the correctly specified model, after which it should steeply decrease to 0 for the misspecified models.
As the approximate model fit indices showed similar computational issues for the CFA-Complex model as for the CFA-Simple model, I first examined to what extent the approximate model fit indices were equivalent (and equal to 1 for BCFI/BTLI or 0 for BRMSEA) across model specifications. The proportions of replications that resulted in equivalent fit index values across model specifications are reported in Table 3. From this Table, it becomes clear that the issue of equivalent approximate model fit indices was less severe than it was for the CFA-Simple model. Equivalent approximate model fit indices were unlikely with complete data or with larger sample sizes. However, the approximate model fit indices were much more likely to be equivalent if 50% of values were missing in 9 items and the sample size was relatively small (i.e., n = 50 or 100).

The model selection proportions for the approximate fit indices reported in Figure 24 are based solely on replications for which the fit indices were different across model specifications. That means that for the right-most column in the figure, displaying the conditions with the largest number of missing values, the points and lines are based on just a few replications, particularly for n = 50 and 100. Thus, those results should be interpreted with caution.

Table 3. CFA-Complex: Proportion of replications for which the approximate model fit indices were equal across all three model specifications.

Sample Size	Complete Data	Number of items and percentage MAR			
		3 items		9 items	
		15%	50%	15%	50%
50	0.006	0.043	0.362	0.311	1.000
100	0.008	0.026	0.228	0.203	0.997
250	0.000	0.000	0.006	0.003	0.806
500	0.000	0.000	0.000	0.000	0.150

Overall, all fit indices were likely to select the correctly specified model for $n = 250$ and 500. For $n = 100$ and $n = 50$, the performance of the fit indices diverged somewhat more clearly. Overall, the PPP-value was least likely to select the correctly specified model, particularly for $n = 50$ or when data contained missing values. However, the PPP-value still selected the correctly specified model most of the time across sample size and missing value conditions. The approximate fit indices were most likely to select the correctly specified model across conditions when $n = 50$ or 100. It should be noted that these indices are not included in the plot for $n = 50$ with 50% missing values in 9 items (top-right plot) as their values were equal to 1 (or 0 for the BRMSEA) across all three model specifications (Table 3). Thus, as long as the approximate fit indices are properly estimated, they will most likely select the correctly specified model across all included conditions.

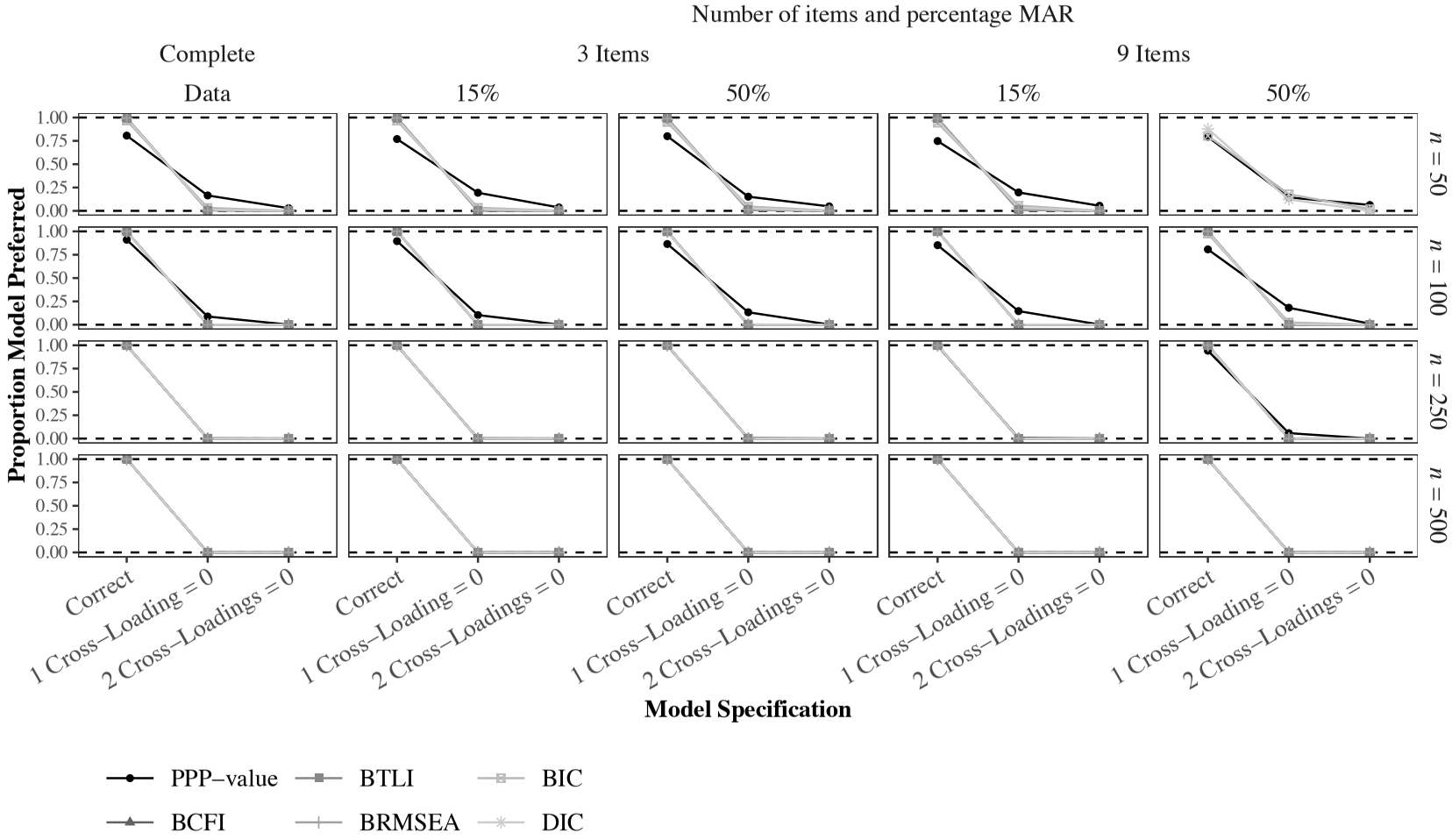


Figure 24. CFA-Complex: Proportion of times a model was selected based on each fit index's value across simulation conditions.

3.3.4 LGM

In this section, I will present the results for the LGM population model. For this population model, I included the prior specification as a factor in the simulation, examining the impact of specifying diffuse, aligned, or diverging priors.

3.3.4.1 The Value of the Model Fit Indices

Model fit indices should increasingly indicate that a model fits the data poorly as model misspecification becomes more severe. As with the CFA-Simple and CFA-Complex population models, I used boxplots to assess whether the model fit indices followed this pattern for the CFA-Complex population model. However, the organization of these plots has changed to accommodate the prior specification conditions. The *x*-axis now reflects the prior specification. Within each prior specification, the grouped boxplots represent the four model specification levels.

Results for the PPP-value appear in Figure 25. Each plot includes a horizontal line at PPP = .05 to emphasize at what point a model might be rejected based on this cutoff value. Across all conditions, the PPP-value decreased as model misspecification became more severe. However, the prior specification greatly affected the PPP-value. Whereas the aligned and diffuse prior specifications followed a similar pattern, if diverging priors were specified, the PPP-value was visibly lower for the correctly specified model. This effect was more substantial for $n = 50$ and 100 and conditions with a large number of missing values. However, even for $n = 500$, the PPP-value was still lower if diverging priors were specified compared to diffuse or aligned priors. Further, the sample size itself greatly affected to what extend the PPP-value decreased. Focusing on the diffuse prior specification, for $n = 50$, the PPP-value was unlikely to drop below .05, unless the quadratic slope was omitted entirely. In contrast, for the largest sample size ($n = 500$), the median PPP-value dropped below .05 if the quadratic slope variance was constrained to 0. Missing values reduced the ability of the PPP-value to differentiate between different levels of model misspecification. For example, the PPP-value decreased to a similar extent for complete data with a sample size of 50 as it did for data with 50% missing values in nine items with a sample size of 250.

Results for the BCFI and BTLI appear in Figures 26 and 27. A horizontal line in each plot emphasizes the point at which a model might be rejected based on the cutoff value. As both indices follow a similar pattern, I will discuss only the BCFI (Figure 26) here. In this figure, we see that, although the BCFI decreased as model misspecification became more severe, its value was not as sensitive to model misspecification as the PPP-value. However, the BCFI did not appear to be as sensitive to the prior specification as the PPP-value, although the BCFI was slightly lower and varied more across replications with diverging priors compared to diffuse or aligned priors. The BCFI increased and varied less as the sample size increased (moving down a column). The BCFI also increased somewhat, and varied more, as the number of missing values increased (moving left to right within a row). The computational issues observed with the CFAs for small samples with missing values appeared less common with the LGM, especially with diverging priors. This is not surprising, as the LGM includes fewer observed variables than the CFA models, resulting in a lower p^* (135 for the CFA-Simple vs. 20 for the LGM).

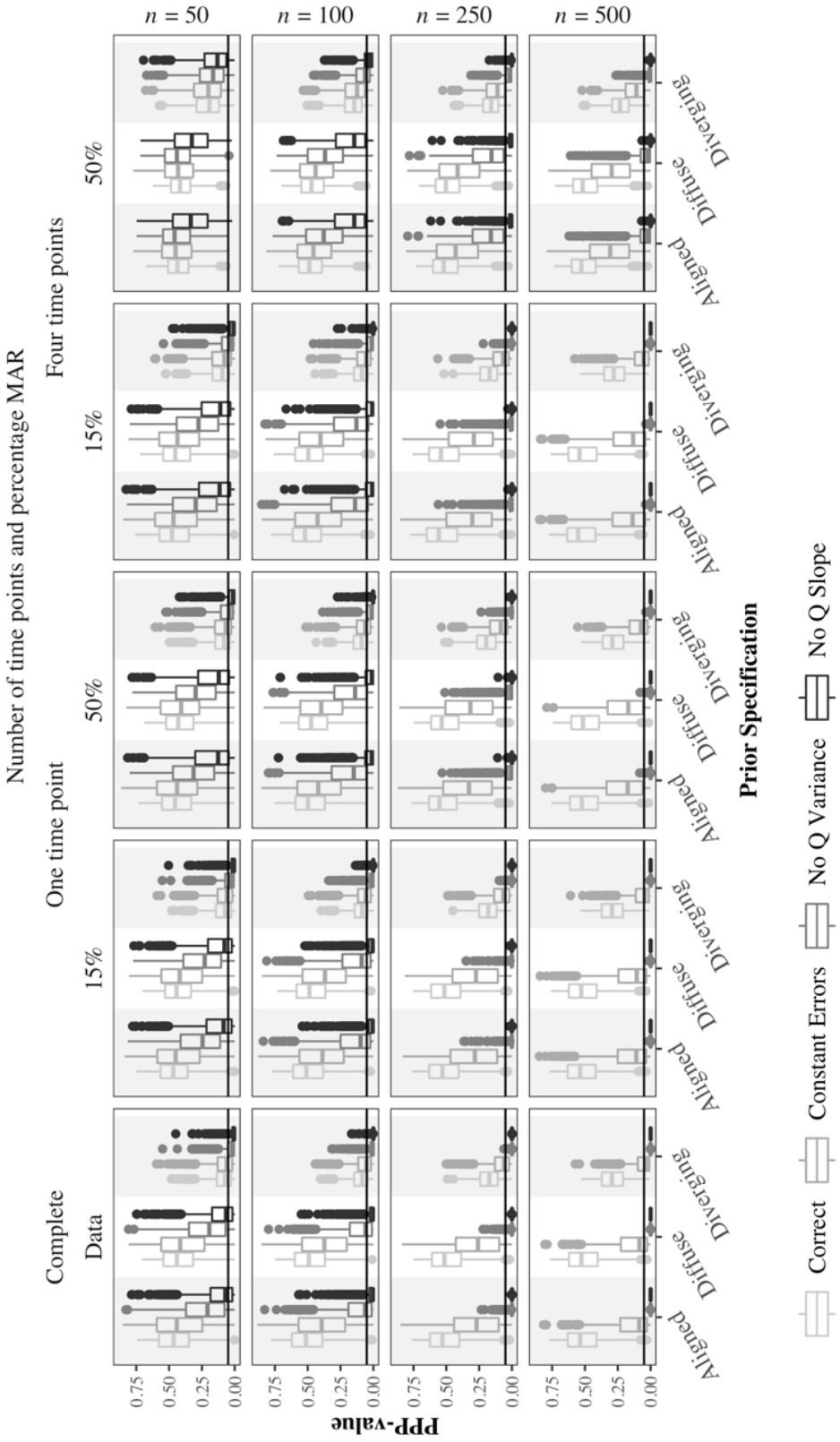


Figure 25. LGM: PPP-value across simulation conditions.

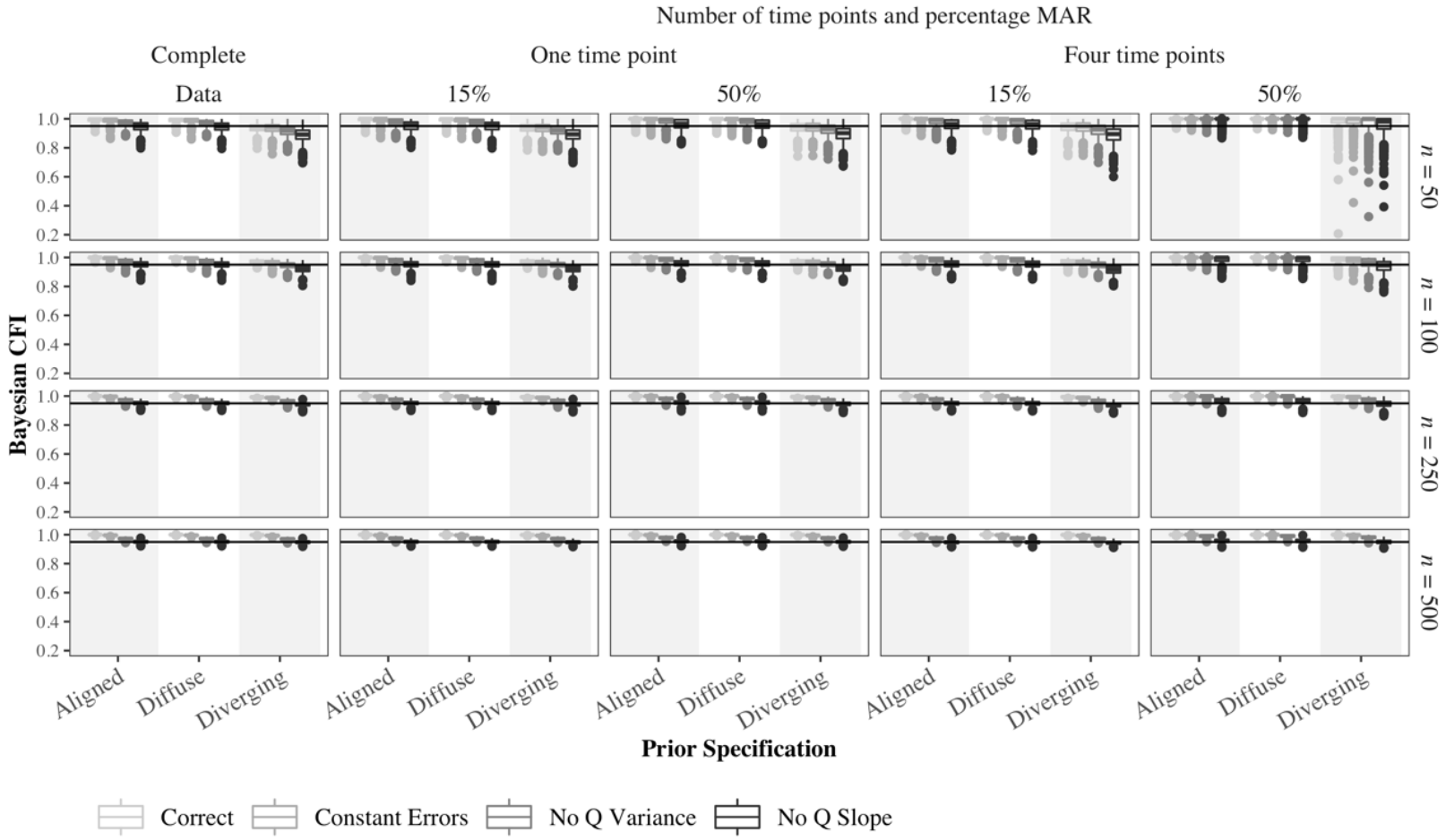


Figure 26. LGM: BCFI across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

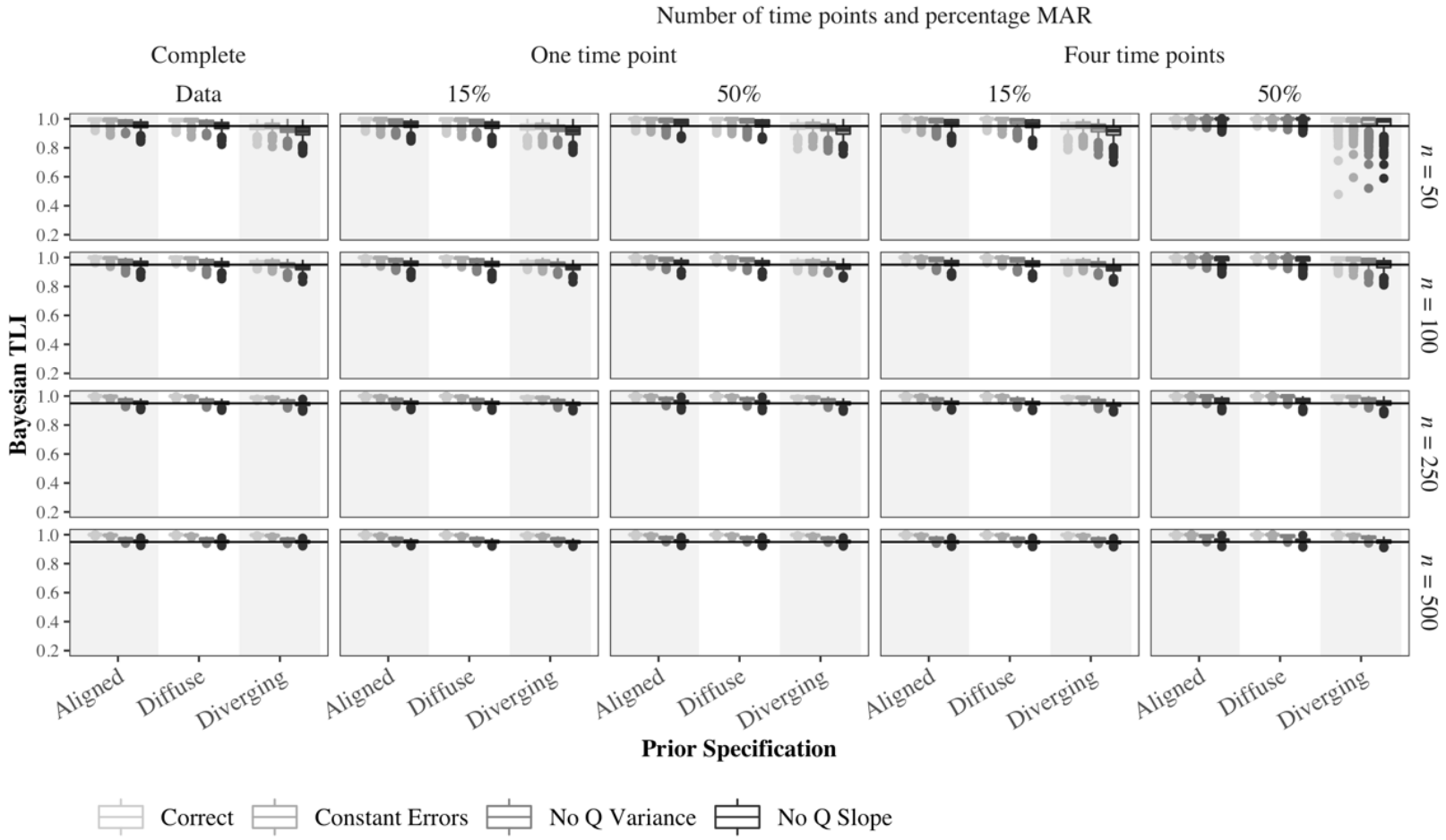


Figure 27. LGM: BTLI across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

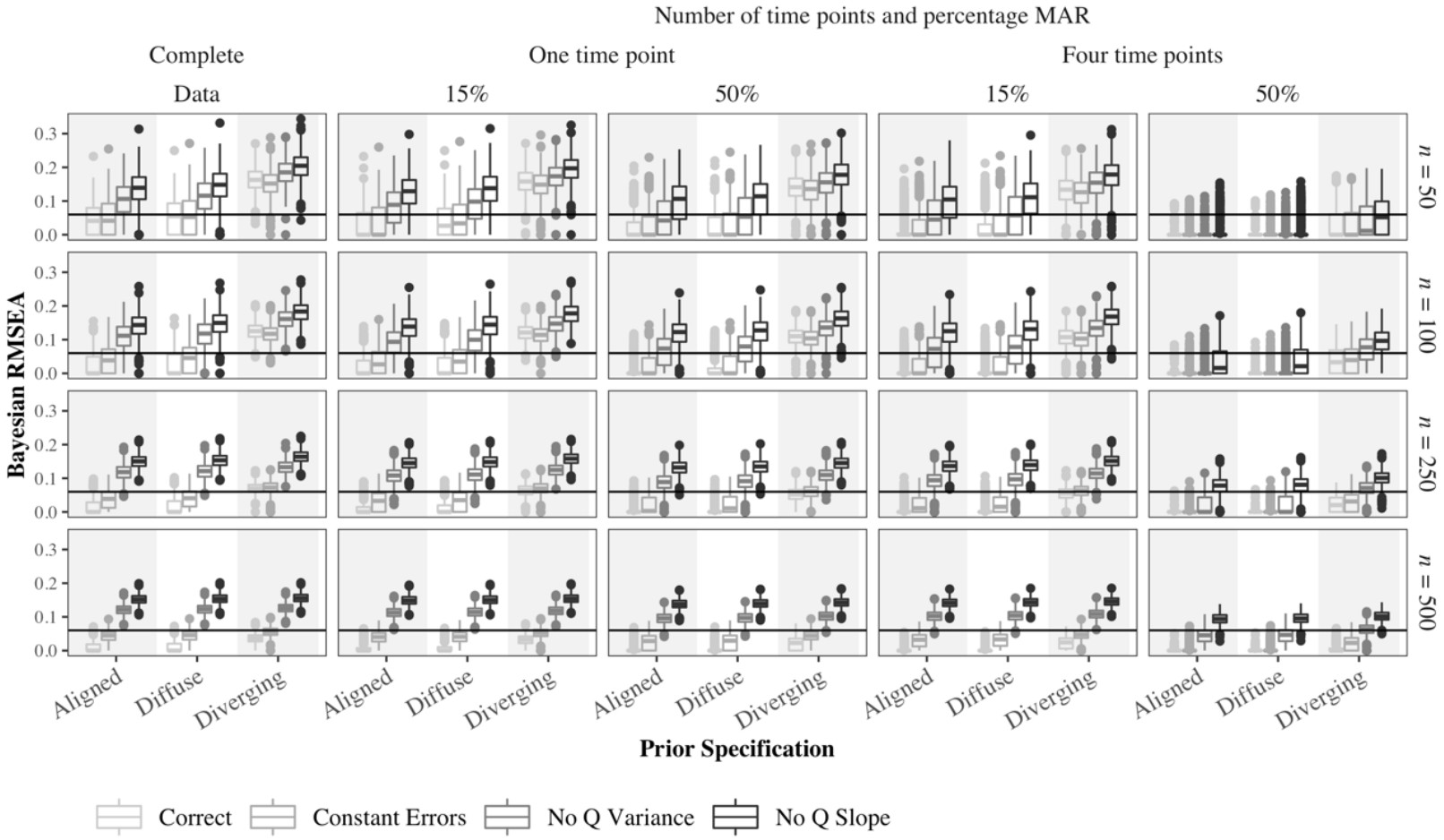


Figure 28. LGM: BRMSEA across simulation conditions. Note that the *y*-axis does not cover the full possible range to highlight subtle changes.

Figure 28 depicts results for the BRMSEA. A horizontal line at .06 is included in each plot to emphasizes the point at which a model might be rejected based on this cutoff value. Compared to the BCFI and BTLI, the BRMSEA was less sensitive to the sample size but more sensitive to the prior specification. The BRMSEA was higher across model specifications for diverging priors compared to aligned or diffuse priors. The impact of the prior specification became smaller as the sample size increased, barely affecting the BRMSEA value for n = 500. The BRMSEA was relatively unaffected by the number of missing values, unless 50% of values with missing at four time points. Under that condition, the BRMSEA was lower, especially for n = 50 and 100. However, even for n = 500, the median BRMSEA was visibly lower across model specification levels compared to conditions with fewer missing values.

3.3.4.2 Does the Model fit the Data well?

An important question for researchers is whether a particular model fits the data well. The PPP-value, BCFI, BTLI, and BRMSEA can all be used to answer this question by using cutoff values or (for the approximate fit indices) using a credible interval. I will first go over the results for the LGM based on cutoff values, after which I will present the result based on credible intervals.

3.3.4.2.1 Using Cutoff Values

Figure 29 and 30 show the proportion of replications that are rejected based on a given fit index's cutoff value. These figures follow a similar layout as Figure 20. However, to examine the impact of the prior specification, the results appear for n = 50 and 100 (Figure 29) and n = 250 and 500 (Figure 30). Ideally, each line starts at 0 for the correctly specified model, after which it should steeply increase to 1 for the misspecified models.

In general, the fit indices rejected misspecified models more often as the sample size increased. Model rejection rates were relatively stable across missing data conditions, unless the sample size was 50 or for the most extreme missing data condition (right-most column in Figures 29 and 30). For n = 50 (Figure 29), rejection rates decreased as the number of missing values increased. Further, if 50% of values were missing in four time points, all fit indices were less likely to reject misspecified models (compared to conditions with fewer missing values). This effect diminished as the sample size increased. Although the values presented in Figures 26-28 suggested that the computational issues with small samples and missing data were not as apparent with the LGM, Figure 29 reveals that the issue may persist even though this is a simpler model.

For the smaller sample size levels (Figure 29), the rejection rates were similar for the aligned and diffuse prior specifications. However, model rejection rates of the correctly specified model were inflated if diverging priors were specified. With diverging priors, the BRMEA rejection rate was close to 1 even if the model was correctly specified. With aligned or diffuse priors, the BRMSEA had the highest model rejection rates across model specification levels, followed by the PPP-value. The BCFI and BTLI rejection rates were similar to that of the PPP-value for n = 50, but lower for n = 100. While the BRMSEA rejection rate was highest for misspecified models, it was also inflated for the correctly specified model, particularly for n = 50. Overall, for smaller sample sizes, with

diffuse or aligned priors, only the BRMSEA (followed by the PPP-value) was likely to reject the most severely misspecified model.

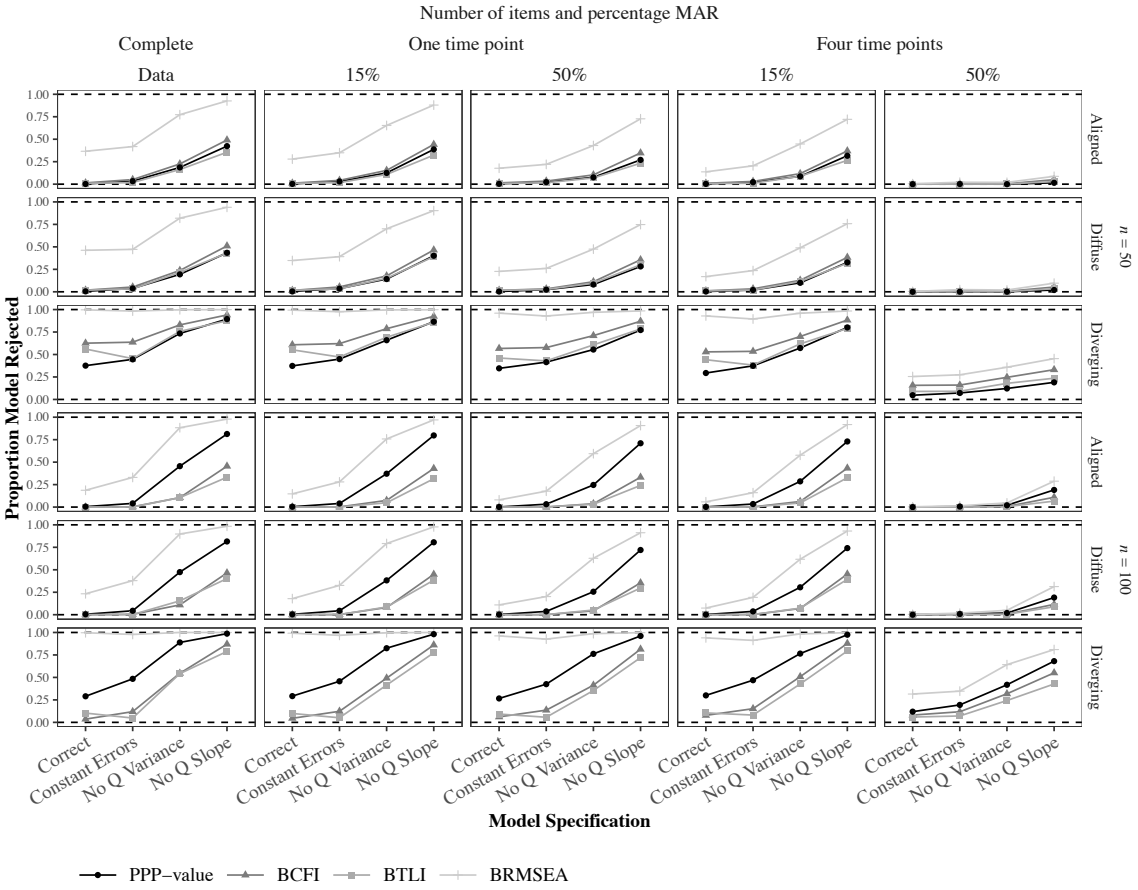


Figure 29. LGM: Proportion of times a model was rejected based on each fit index's cutoff value across simulation conditions for n = 50 and 100.

For the larger sample size levels (Figure 30), the rejection rates were similar for the aligned and diffuse prior specifications. Overall, the impact of diverging priors was diminished for these larger sample sizes. However, for n = 250, BRMSEA model rejection rates of the correctly specified model were still inflated if diverging priors were specified. With aligned or diffuse priors, the PPP-value and the BRMSEA had the highest model rejection rates for misspecified models. The model rejection rates of the BCFI and BTLI were close to zero for all but the most severely misspecified model. Even for that model, the rejection rates of the BCFI and BTLI were visibly lower than the BRMSEA and PPP-value. Overall, for larger sample sizes, the BRMSEA and PPP-value were likely to reject a model that constrained the quadratic slope variance to 0 or omitted the quadratic slope entirely. The model in which measurement errors were constrained to be equal was unlikely to be rejected by these indices across sample sizes, although the rejection rate for this model misspecification slowly increased as the sample size increased.

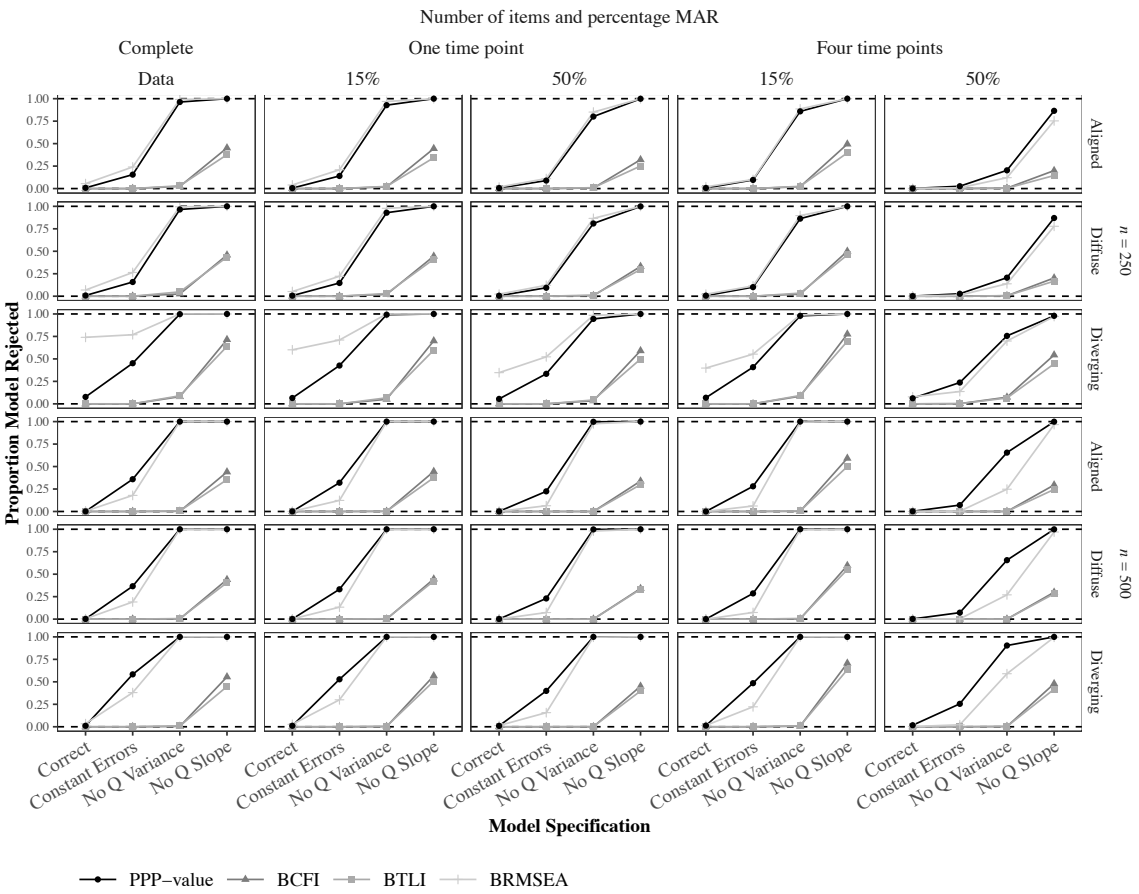


Figure 30. LGM: Proportion of times a model was rejected based on each fit index's cutoff value across simulation conditions for *n* = 250 and 500.

3.3.4.2.2 Using 90% Credible Intervals

For the approximate fit indices, it is possible to use their 90% credible interval to assess model fit. This approach results in one of three conclusions: (1) the model fits the data well (the entire 90% credible interval is beyond the cutoff that denotes good fit), (2) model fit is inconclusive (the cutoff values is within the 90% credible interval), (3) the model does not fit the data (the entire 90% credible interval is < .95 for the BCFI/BTLI or > .06 for the BRSMEA). The results for the BCFI, BTLI, and BRMSEA based on this approach are shown in Figures 31-36 and are organized in the same manner as Figures 21-23, but with separate figures for the smaller (*n* = 50 and 100) and larger (*n* = 250 and 500) sample sizes. Ideally, the entire bar is the lightest shade for the correctly specified model and the darkest shade for the misspecified models. As the pattern of results is similar for the BCFI and BTLI, only the results for the BCFI (Figures 31 and 32) are discussed in detail. The results for the BRMSEA (Figures 35 and 36) will be discussed separately.

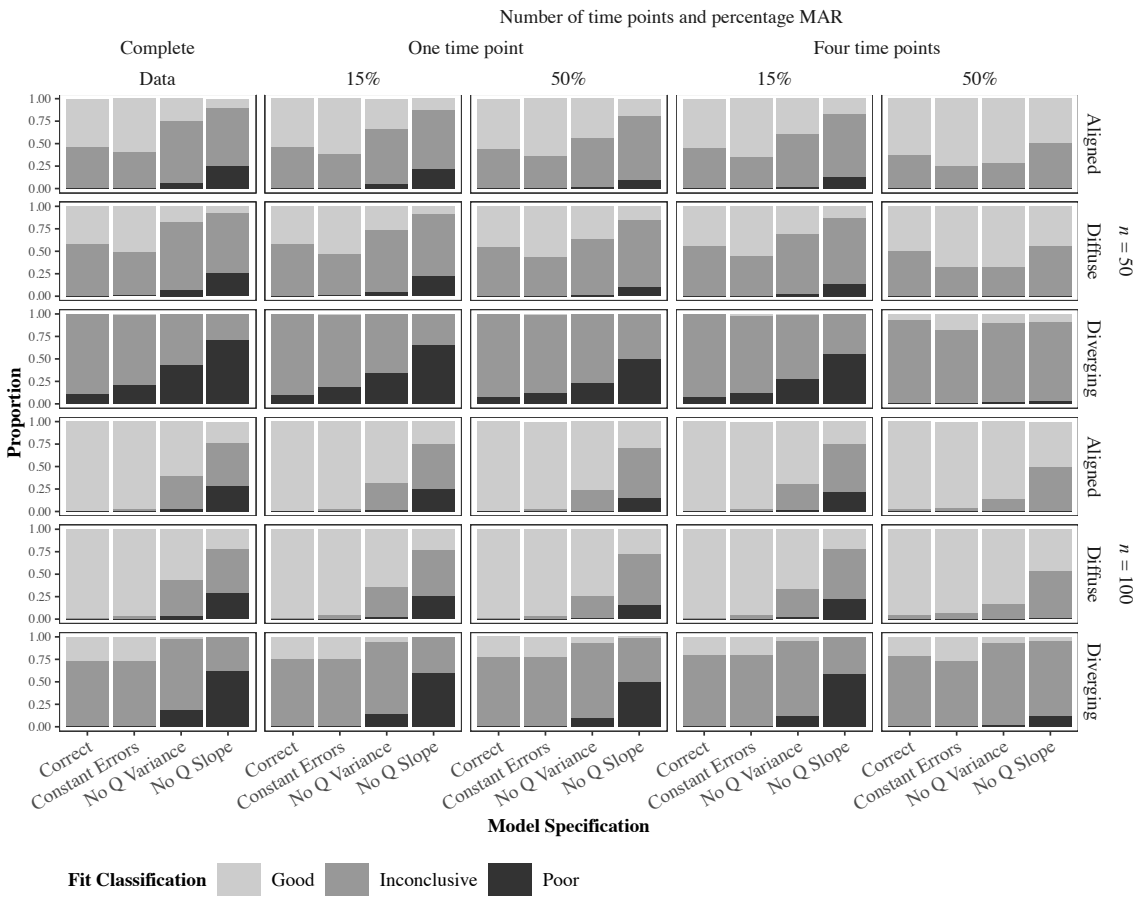


Figure 31. LGM: Model fit classification based on 90% BCFI credible interval for n = 50 and 100.

Overall, the credible intervals were somewhat more likely to indicate poor model fit as the severity of the model misspecification increased. However, as the sample size increased, the credible intervals became more likely to indicate inconclusive or even good model fit for all but the most severe level of model misspecification. For n = 50 and 100 (Figure 31), the use of diverging priors increased the proportion of replications that resulted in inconclusive and, to a lesser extent, poor model fit. The difference between diverging and aligned or diffuse priors was smaller for n = 250 and nonexistent for n = 500 (Figure 32). The presence of missing values increased the probability that the credible intervals would indicate good or inconclusive model fit. If 50% of values were missing in 4 time points, the credible intervals were unlikely to indicate poor or inconclusive model fit (except for the most severe level of misspecification).

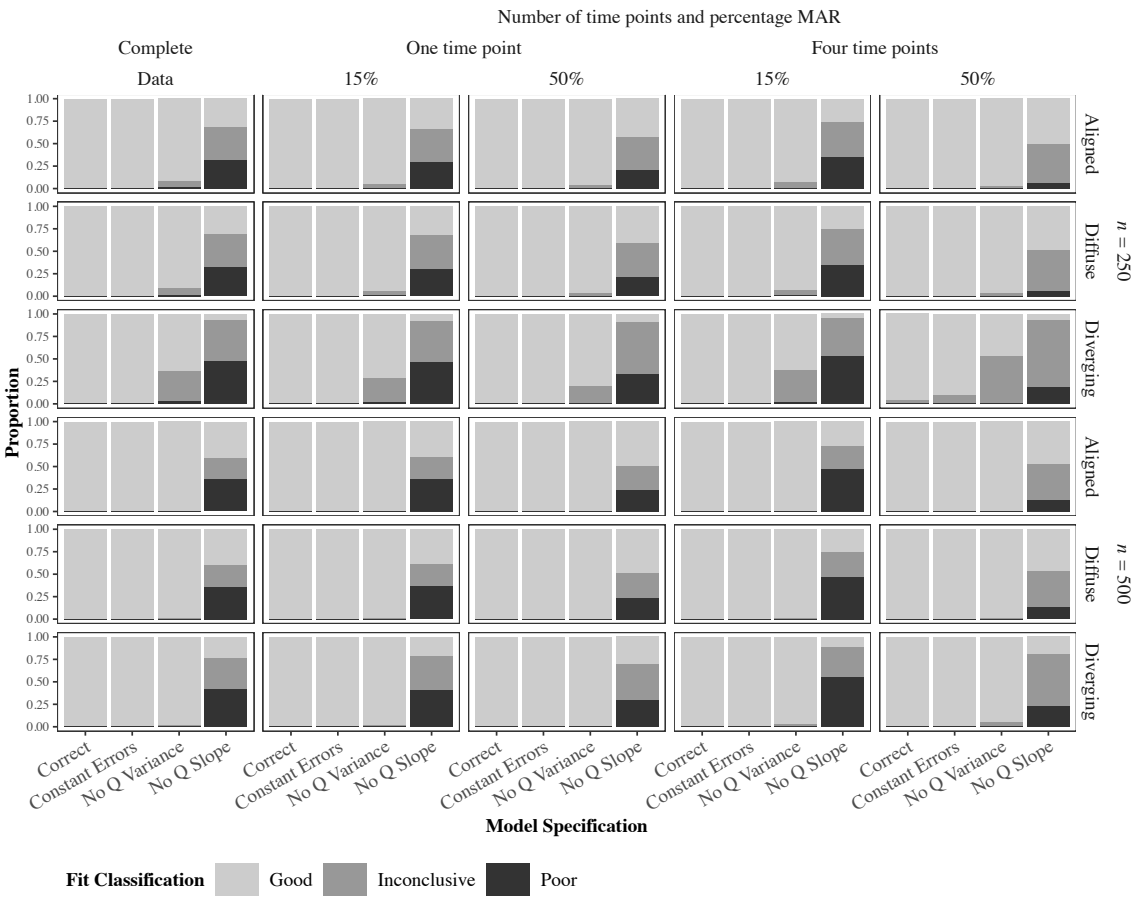


Figure 32. LGM: Model fit classification based on 90% BCFI credible interval for n = 250 and 500.

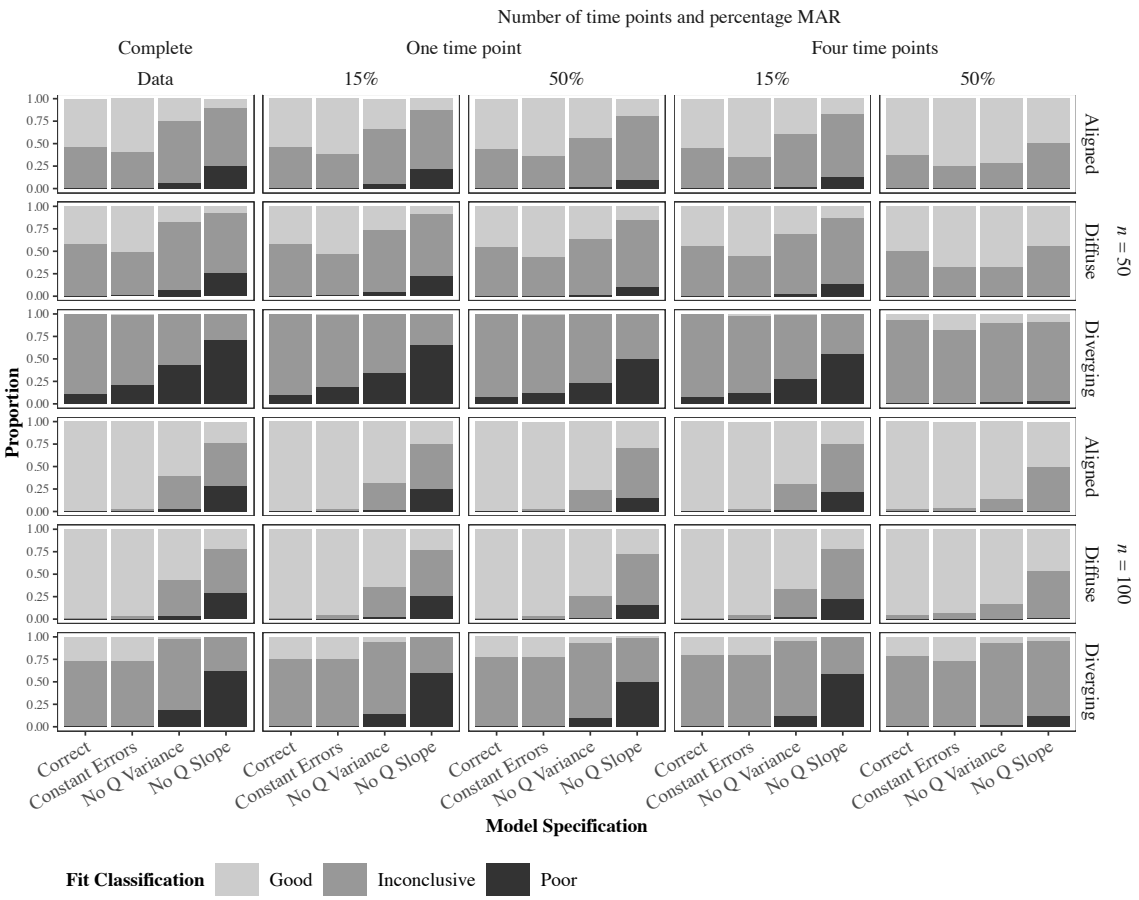


Figure 33. LGM: Model fit classification based on 90% BTLI credible interval for $n = 50$ and 100.

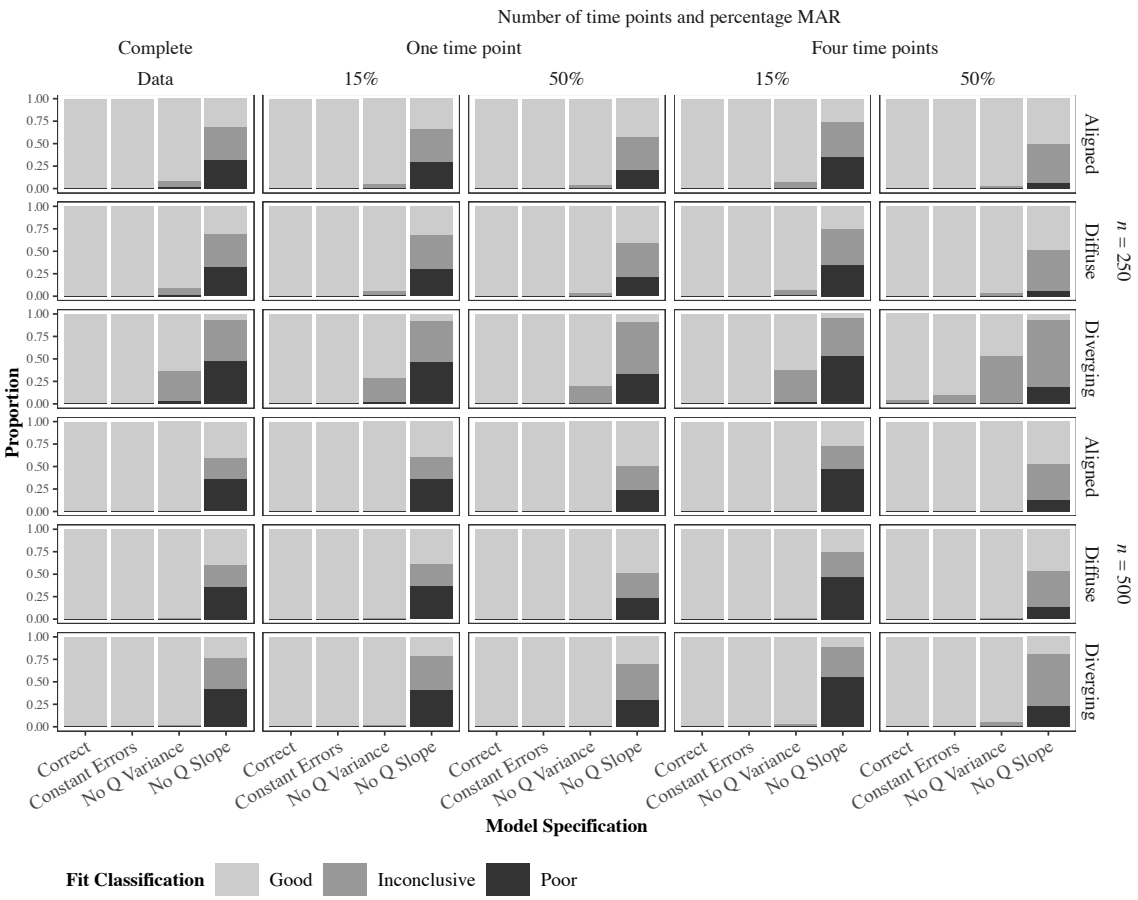


Figure 34. LGM: Model fit classification based on 90% BTLI credible interval for n = 250 and 500.

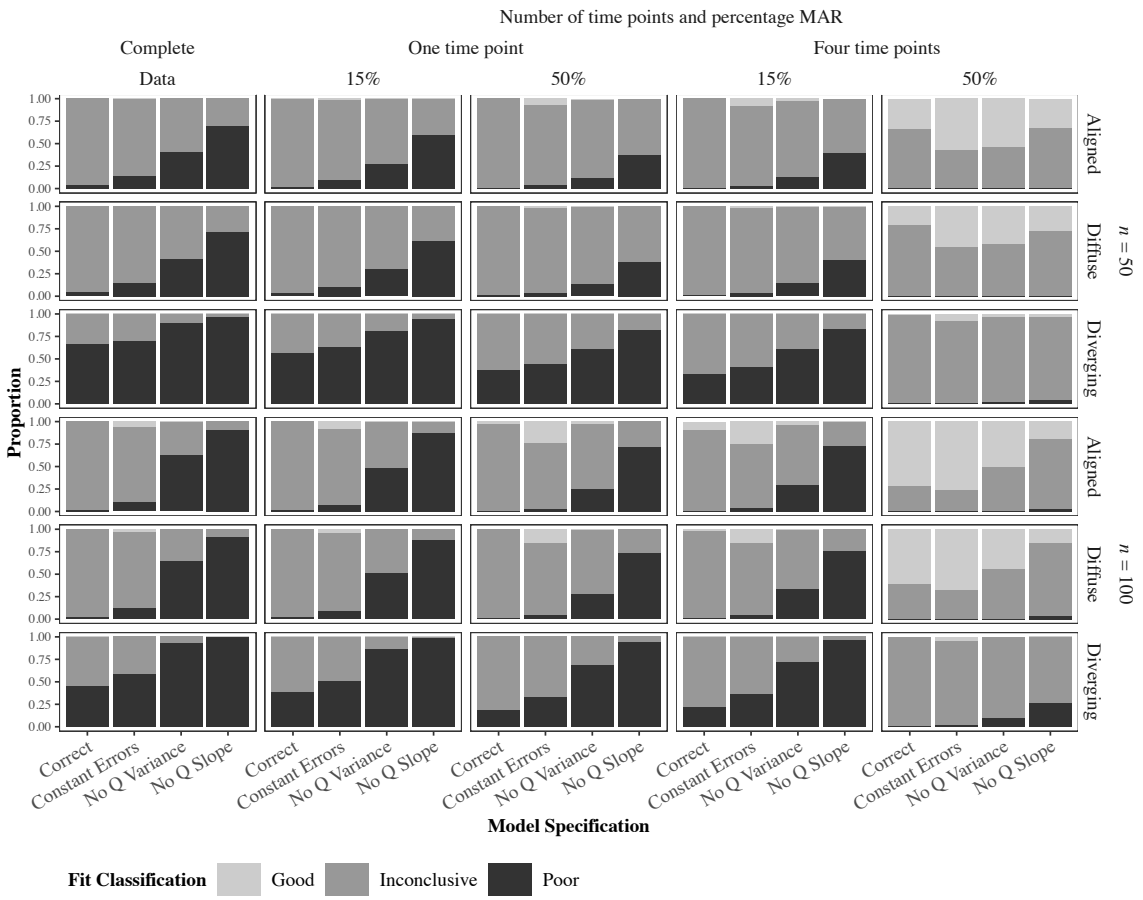


Figure 35. LGM: Model fit classification based on 90% BRMSEA credible interval for *n* = 50 and 100.

For the BRMSEA, the pattern of results was quite different from the BCFI. Overall, the credible intervals were more likely to indicate poor model fit for the two most severe levels of model misspecification as the sample size increased. However, the credible intervals were also likely to indicate inconclusive model fit for correctly specified models, unless *n* = 500 and diffuse or aligned priors are specified (Figure 36). For *n* = 50 and 100 (Figure 35), the use of diverging priors increased the proportion of replications that resulted in inconclusive and poor model fit. Even with larger sample sizes (*n* = 250 and 500; Figure 36), the use of diverging priors resulted in inconclusive BRMSEA credible intervals. With informed or diffuse priors, the presence of missing values increased the probability that the credible intervals would indicate good or inconclusive model fit. If 50% of values were missing in 4 time points, the credible intervals were unlikely to indicate poor or inconclusive model fit (except for the most severe level of misspecification).

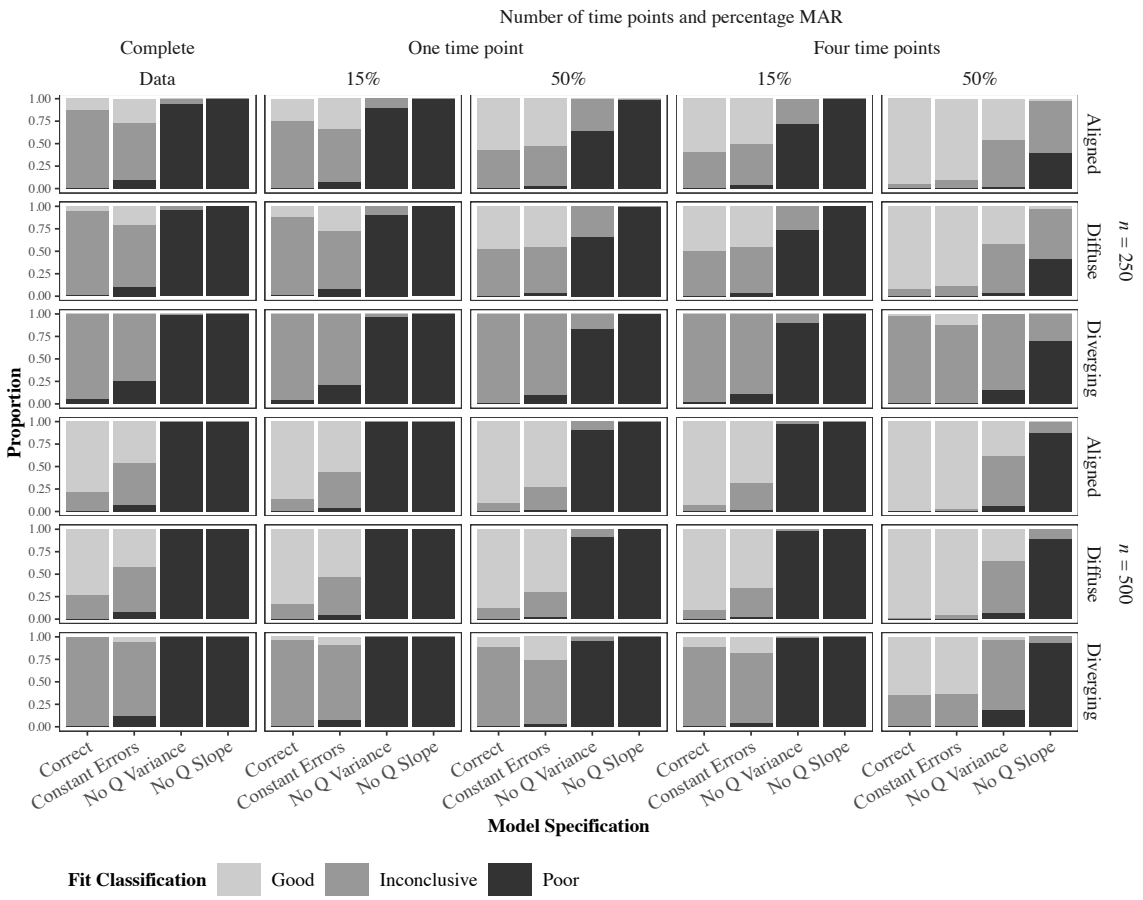


Figure 36. LGM: Model fit classification based on 90% BRMSEA credible interval for *n* = 250 and 500.

3.3.4.3 Model Selection

Model fit and selection indices can also be used for model selection. Figures 38 and 39 shows the proportions of replications selected based on a given fit index's value. These figures follows the same layout as Figures 29 and 30. Ideally, each line starts at 1 for the correctly specified model, after which it should steeply decrease to 0 for the misspecified models.

Although the approximate model fit indices did not appear to show similar computational issues as for the CFA-Simple or CFA-Complex model, I still examined to what extent the approximate model fit indices were equivalent (and equal to 1 for BCFI/BTLI or 0 for BRMSEA) across model specifications. The proportions of replications that resulted in equivalent fit index values across model specifications are reported in Figure 37. This figure is more complex compared to Tables 2 and 3 for several reasons. Whereas the approximate model fit indices were always equivalent across all model specifications for the CFA models, that was not the case for the LGM. Within each plot in the figure, I have reported the proportion of replications that resulted in different or equal approximate fit index values across two, three, or all four

specifications. Further, the proportions are reported for each sample size and prior specification separately.

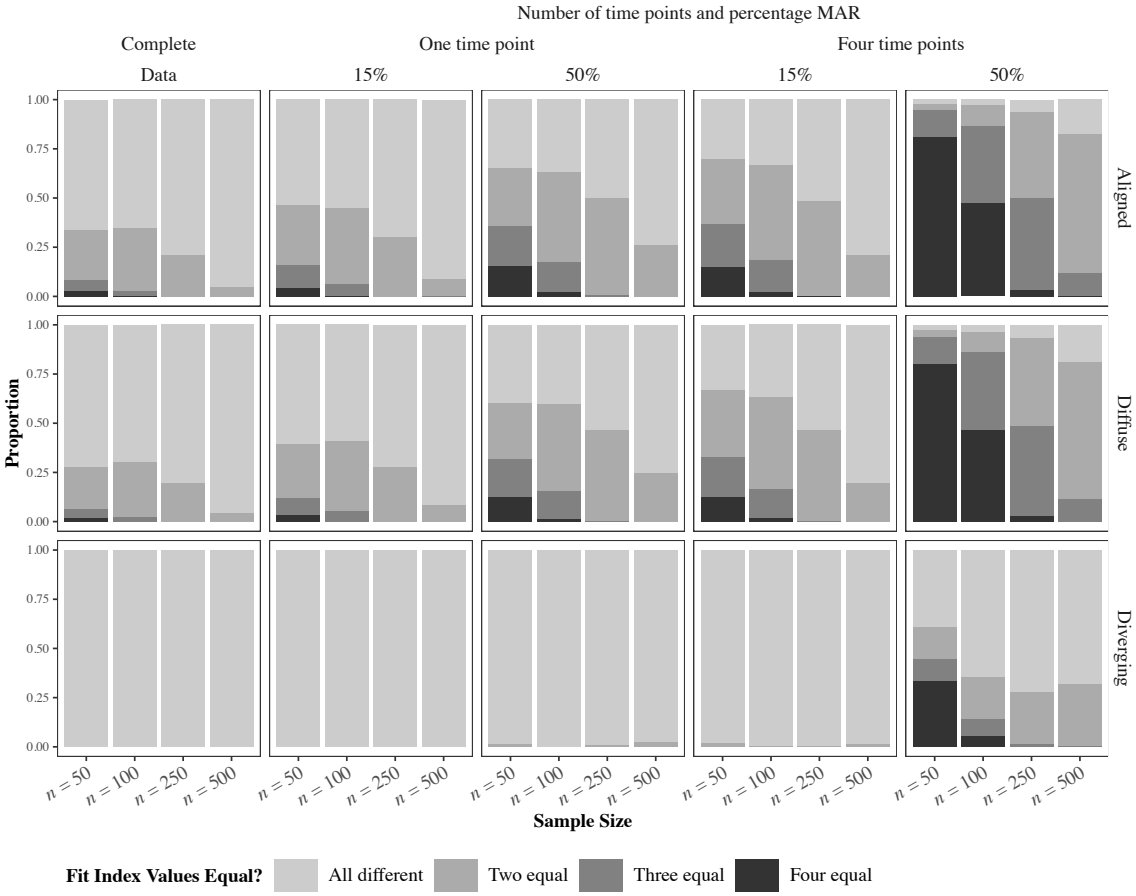


Figure 37. LGM: Proportion of replications for which the approximate model fit indices were different or equal across two, three, or four model specifications.

From this figure, several things become clear. First, equivalent model fit index values were much less common when diverging priors were specified. That is likely because the diverging priors caused the approximate model fit index values to reflect worse fit across specifications, making it less likely that their value was so close to perfect fit that the index value could not be computed at all. From this we learn that if the discrepancy between the baseline and estimated model becomes too small (indicating poor fit), the computational issues appear less likely to arise than if the discrepancy is larger (indicating good fit). Second, with complete data, model fit indices were more likely to be different or equivalent across two model specifications (i.e., the correctly specified model and the model with constant measurement errors) than across three or all four model specifications. That is encouraging, as it means that the approximate model indices were still able to differentiate between substantively irrelevant and relevant model misspecification. However, as the number of missing values increased, the proportion of replications that resulted in equivalent model fit index values across all four model specifications increased and became more likely. Third, although the proportion of

replications that resulted in equivalent model fit index values was never 1 (i.e., all replications), the overall pattern of observing any number of equivalent model fit index values was somewhat between that of the CFA-Simple and CFA-Complex. Even with n = 500 and with complete data, some replications still resulted in two equivalent model fit index values. However, three or four equivalent model fit index values did not occur for n = 250 and 500 with complete data.

The model selection proportions for the approximate fit indices reported in Figures 38 and 39 are based solely on replications for which the fit indices were different across all model specifications. That means that for the right-most column in the figure, displaying the conditions with the largest number of missing values, the points and lines are based on just a few replications. Thus, those results should be interpreted with caution.

The results displayed in the figures below reveal several interesting patterns. Across sample sizes, the fit indices were most likely to select either the correctly specified model or the model in which the measurement errors were held constant. The likelihood of selecting the correctly specified model increased across the board as the sample size increased. The BIC, and to a lesser extent the DIC, preferred the model with constant measurement errors over the correctly specified model. That pattern likely reflects the preference of the BIC and the DIC for parsimonious models. As the sample size increased, the DIC started to prefer the correctly specified model, whereas the BIC remained more likely to select the model with constant measurement errors. Specifying diverging priors had a less apparent effect on model selection than it had on model rejection rates. If fit indices are compared across multiple specifications, the use of diverging priors did not suddenly increase the likelihood of selecting a severely misspecified model. However, for n = 50, the use of diverging priors did increase the likelihood that the PPP-value, BCFI, BTLI, and BRMSEA selected the model with constant measurement errors. For n = 100, the BTLI and the BRMSEA remained more likely to select this slightly misspecified model. The impact of the diverging priors further diminished with larger sample sizes (Figure 38).

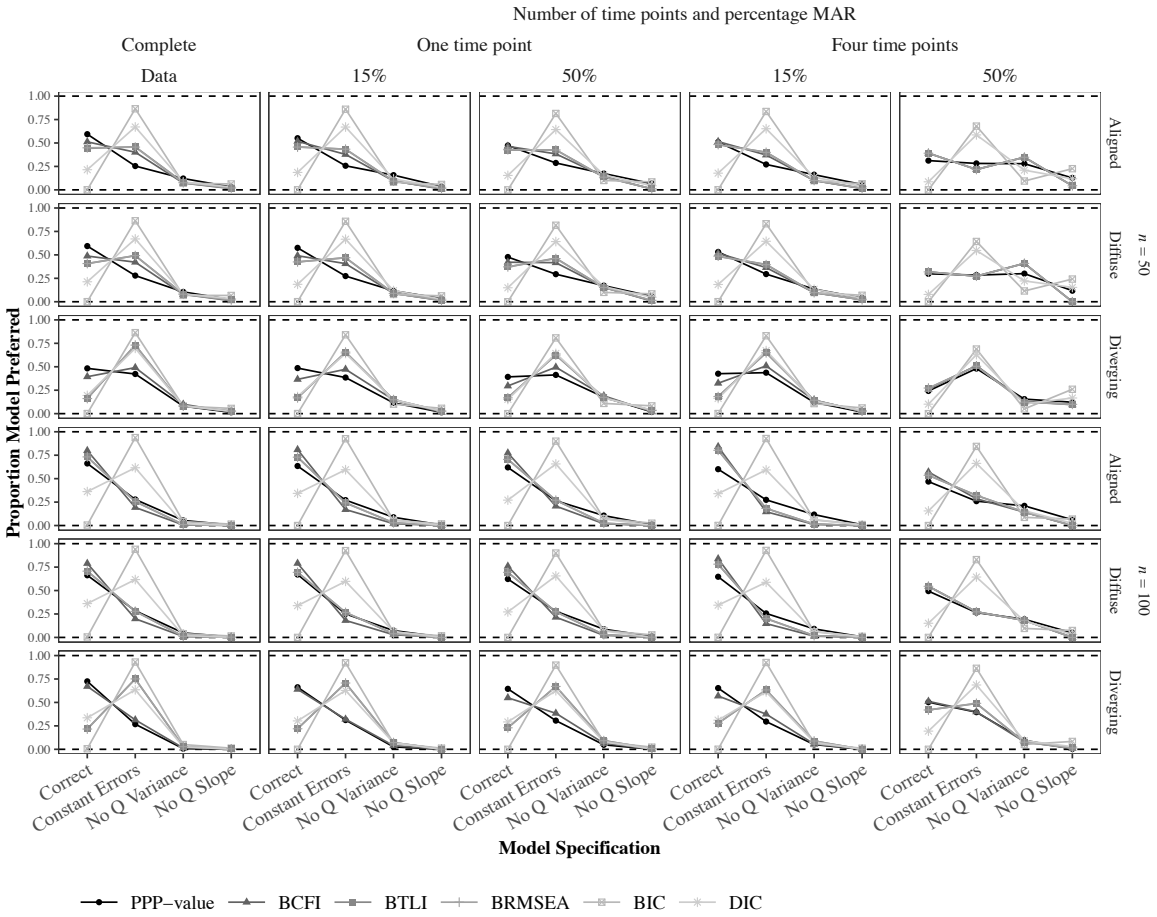


Figure 38. LGM: Proportion of times a model was selected based on each fit index's value across simulation conditions for n = 50 and 100.

The presence of missing values only meaningfully affected the pattern of model selection if 50% of values were missing in four time points (the right-most column in Figures 37 and 38). The impact was especially noticeable for the smallest sample size. With aligned or diffuse priors, the PPP-value, the BCFI, the BTLI, and the BRMSEA were about equally likely to select the correctly specified model as it was to select the first two misspecified models. For n = 50, this level of missing data also increased the likelihood that the BIC and DIC selected the most misspecified model. As the sample size increased, the preference of the DIC for the model with constant measurement errors was more apparent when 50% of values were missing in four time points.

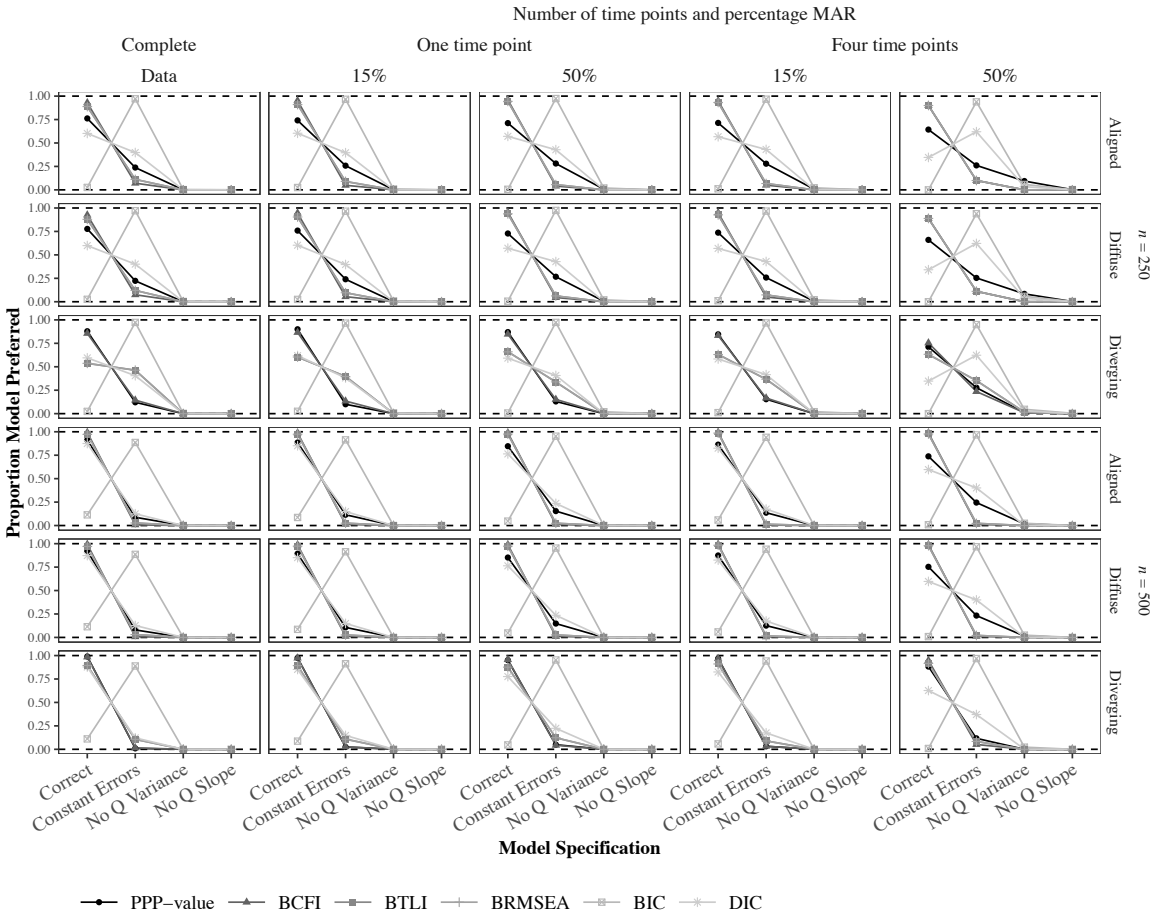


Figure 39. LGM: Proportion of times a model was selected based on each fit index's value across simulation conditions for n = 250 and 500.

3.4 Discussion

Study 1 examined the ability of the M*plus* implementation of the PPP-value, BCFI, BTLI, BRMSEA, BIC, and DIC to detect model misspecification in Bayesian SEM. Gaining a better understanding of the extent to which those indices can be used for model fit and selection purposes will help applied researchers judge the appropriateness of their statistical models. I will first discuss the overall findings regarding model fit assessment and model selection, followed by a discussion of the impact of missing data and the role of priors.

3.4.1 Model Fit Assessment

Overall, it appears that the PPP-value may be the most useful for model fit assessment. Across population models, it was most strongly affected by each level of model misspecification. Furthermore, as the sample size increased, the PPP-value was more likely to reflect poor model fit for misspecified models, even if the severity of the misspecification was relatively small. However, if the sample size is limited (e.g., n =

50), the PPP-value is unlikely to reflect poor model fit if the misspecifications are substantively irrelevant (i.e., for the CFA-Simple).

In contrast, the approximate fit indices were less clearly affected by model misspecification, often indicating good model fit for misspecified models. Further, focusing on the LGM population model, the results of Study 1 show that the Bayesian approximate fit indices, particularly the BCFI and BTLI, were less sensitive to misspecification in the marginal mean structure compared to the PPP-value. This finding is in line with previous research on the frequentist versions of the CFI and TLI (Wu et al., 2009; Wu & West, 2010). In contrast, the BRMSEA was sensitive to misspecification in the marginal mean structure, and overall performed better in terms of model fit assessment than the BCFI and BTLI. Most worrisome was the finding that the approximate fit indices appeared to indicate better model fit across all levels of model (mis)specification as the sample size increased. At $n = 50$, the approximate fit indices were likely to indicate poor model fit for the correctly specified model. However, at $n = 500$, the approximate fit indices were less likely to indicate poor model fit for all but the most severe levels of misspecification. Among the three indices, the BRMSEA appeared somewhat more sensitive to model misspecification compared to the BCFI and the BTLI.

One potential advantage of the Bayesian approximate fit indices is that they possess a posterior distribution that can be used to create a 90% credible interval (different percentages can also be examined). This interval has been suggested to provide a more nuanced reflection of model fit compared to the posterior mean value by itself (Asparouhov & Muthén, 2020). By using the interval, a researcher may conclude that model fit is good, inconclusive, or poor. Based on the results of Study 1, it appears that this method may only be useful for small sample sizes (e.g., $n = 50$). However, this strategy could result in concluding inconclusive model fit even if the correct model was specified. For larger sample sizes, inconclusive model fit occurred less often, but could still be useful for detecting model misspecifications that are missed by relying on the cutoff value. Overall, the conclusion drawn based on the credible interval appeared to rely on multiple factors that were not related to the model misspecification itself, reducing its general utility for applied researchers.

It should be noted that Asparouhov and Muthén (2020) argued that the approximate model fit indices should only be used with larger sample sizes, which they defined as "more than 100 or even 200" (p. 9). However, the results of the current study seem to indicate that the utility of the approximate fit indices at larger sample sizes may be limited. The indices (especially the BCFI and BTLI) are likely to indicate that the model fits the data approximately well, even if the level of misspecification is substantively relevant.

3.4.2 Model Selection

For the CFA-Simple and CFA-Complex, all model fit indices selected the correct model 100% of the time if $n = 500$, and the vast majority of the time if $n = 250$. Even if $n = 50$, the correct model was extremely likely to be selected for the CFA-Complex population model. For the CFA-Simple model, model selection accuracy was lower if $n = 50$, particularly for the DIC. This finding may reflect the DIC's preference for parsimonious models in the presence of substantively irrelevant model misspecification. For the LGM,

a similar pattern emerged for the DIC, and to a lesser extent the BIC. A model in which the measurement errors at each time point were constrained to be equal was preferred by those two indices over the correctly specified model if n = 50 or 100. While the correct model became more likely to be selected with the BIC if n = 250 or 500, model selection based on the DIC kept preferring the model with constrained measurement errors. The correct model was most likely to be selected using the remaining model fit indices, with this preference for the correct model increasing as the sample size increased. These findings, combined with the findings discussed above regarding model fit assessment, underline the importance of comparing multiple models within one study. When multiple models are compared, the fit indices were generally able to select the correct model, or a model with a substantively irrelevant level of misspecification.

Although the approximate model fit indices were not developed for model selection, it appears that they might be more suitable for this purpose, under the conditions examined in the current study. In the current study, a difference of any size across model specification was used for model selection. However, the frequentist literature has pointed out that an alternative degree of change may need to be used instead (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2005; Rutkowski & Svetina, 2014). Future research could investigate whether a specific amount of change in the value of the BCFI, BTLI, and BRMSEA can be used for model selection, or if such guidelines depend on factors such as the sample size and model type.

One critical limitation to the use of the approximate fit indices for model selection emerged from the current study. As the number of missing values increased, the approximate fit indices were more likely to reflect "perfect" fit (i.e., 1 for BCFI/BTLI or 0 for BRMSEA) across model specifications. This issue occurred for all three population models investigated, although it appeared less prevalent for the CFA-Complex model. Furthermore, model fit index values were less likely to be equivalent if diverging priors were specified. However, the use of diverging priors may still be problematic for model fit and selection, as further discussed in Section 3.4.4. If a fit index value is equivalent across model specifications, it cannot be used for model selection. Thus, researchers may need to rely on other fit indices for model selection (and model fit assessment) if missing values are prevalent in their data.

In addition to these general patterns of performance of the model fit indices, the results of Study 1 also showed that the ability of the model fit indices to select the correct model depended on the presence of missing data and the prior specification. These two factors will be discussed in more detail in the following two sections.

3.4.3 The Impact of Missing Data

The results of Study 1 indicate that missing data have a different effect on each of the model fit indices. Starting with model fit assessment, an increasing number of missing values consistently reduced the ability of the PPP-value to detect model misspecification. The PPP-value was centered around 0.5 (indicating good model fit) for the correctly specified model independent of the number of missing values. However, as the number of missing values increased, the PPP-value declined to a lesser extent for misspecified models. The PPP-value performed about equally well for a sample of n = 50 without missing values as it did for a sample of n = 250 with 50% missing values in the majority

of the observed variables. This relatively straightforward pattern mirrors results found in previous research on this implementation of the PPP-value (Asparouhov & Muthén, 2020).

In contrast, the pattern of findings for the approximate model fit indices was not as straightforward. With increasing numbers of missing values, the approximate model fit indices indicated better model fit for the correctly specified model. In addition, as the number of missing values increased, the approximate model fit indices decreased (or for BRMSEA: increased) less as the level of model misspecification became more severe. This pattern of results extends previous research on the approximate fit indices with missing data (Asparouhov & Muthén, 2020). The current study design differed from the previous design in terms of the sample sizes included. Asparouhov and Muthén (2020) generated missing values only for samples of n = 300, 1000, and 5000. The current study extends their findings by demonstrating that the use of approximate fit indices as model fit indices (as opposed to model selection indices) is likely to result in falsely concluding good model fit if the data contain missing values.

Furthermore, for the CFA models (and to a lesser extend the LGM), the approximate fit indices could not be computed for the smallest sample size (n = 50) with the largest number of missing values (50% in 9 out of 15 variables). As was discussed in the Results, it appears that the approximate fit indices cannot be computed for situations in which the number of observed variables is relatively large compared to the number of observed values as this reduces the difference between the baseline and estimated model discrepancy too much. It should be noted that this problem did not arise to the same extent for n = 50 with complete data. Thus, it is not just the overall sample size that makes the difference, but the sample size for each observed variable. This computational issue did not arise for the PPP-value, highlighting the PPP-value as a valuable tool for model fit assessment for small samples with missing data.

Missing values had a smaller effect on the ability of the model fit indices to select the correct model out of a set of estimated models. For all indices, model selection performance was not meaningfully affected by missing values unless 50% of values were missing in a majority of the variables. However, at that level of missing data, the fit indices were less likely to select the correct model for the CFA-Simple and the LGM, but not for the CFA-Complex. It may be that if the model misspecifications were substantively irrelevant, as they were for the CFA-Simple, the presence of missing values further muddles the difference in model fit across different (mis)specifications. For the LGM, it is less clear why a large number of missing values had such an impact on the performance of the fit indices to select the correct model. It may be related to the smaller number of parameters or observed variables that are involved in estimating the LGM as compared to the CFA-Complex model. Alternatively, it may be due to the way missing data were generated following a dropout pattern for the LGM with missing values in multiple observed variables. When 50% of values were missing across 4 time points, that meant that 80% of values were missing at the final time point. This sparsity of data may seriously reduce the ability to differentiate between models with and without a quadratic slope (or quadratic slope variance). Increasing the overall sample size reduces the negative impact of missing data on the performance of the fit indices in terms of model selection.

Finally, for the CFA-complex model, it appeared that there was no negative effect of missing values in variables that were involved in the model misspecification. This factor was examined in the current study as previous research had indicated that the frequentist RMSEA and CFI might indicate better model fit when the missing data are located in the part of the model with a misspecification (Zhang & Savalei, 2020). The explanation for this discrepancy may lie in how missing values were treated: previous research looked at ML estimation, which used full information maximum likelihood (FIML) to address missing data. As discussed in the introduction, Bayesian estimation relies on DA, in which values for the missing data are predicted based on the model, the current parameter estimates, and the observed data at each iteration of the MCMC chain. It appears that the location of missing values relative to the location of the model misspecification is less critical when Bayesian estimation is used.

3.4.4 The Role of Priors

The ability to incorporate prior knowledge into the analysis through Bayesian estimation is often lauded as a major advantage (e.g., Smid, McNeish, et al., 2019; van de Schoot et al., 2014, 2017). However, diverging priors may negatively affect the ability of model fit indices to select the correctly specified model (Cain & Zhang, 2019; Liang, 2020). The results of Study 1 support the hypothesis that, when it comes to model fit and selection, there may also be downsides to specifying informative prior distributions in Bayesian SEM. For the LGM population model, specifying informative priors that aligned with the population values, as compared to diffuse priors, did not noticeably improve model fit assessment or model selection for any of the included indices. However, specifying informative priors that diverged from the population values confounds the association between model misspecification and the values of model fit indices. Garnier-Villarreal and colleagues (2019) also observed this pattern of results with an empirical example using their implementation of the approximate model fit indices (i.e., BCFI, BTLI, and BRMSEA). The results of Study 1 add to their finding by demonstrating that, in addition to the M*plus* implementation of the approximate fit indices, diverging priors also affected the performance of the PPP-value, the BIC, and the DIC. The impact of diverging priors on the model fit and selection indices was still apparent for $n = 250$. In terms of model fit assessment, the BRMSEA was most affected by diverging priors, followed by the PPP-value, and finally, the BCFI and BTLI. One notable finding is that for $n = 50$ and 100 (and also $n = 250$ and 500 for the BRMSEA), model fit conclusions based on the 90% credible interval were more likely to be inconclusive if diverging priors were specified. Specifying diverging priors had a less detrimental impact on the ability of the fit indices to select the correctly specified model. This finding indicates that, although the values of the fit indices were biased with diverging priors, they still followed a pattern of indicating worsening model fit as model misspecification increased. The only exception occurred with $n = 50$, where the PPP-value and the BCFI had an increased probability of selecting the misspecified model with constant measurement errors. As this misspecification results in a more parsimonious model, and can be considered substantively irrelevant, this may not be a problematic pattern for applied researchers who rely on relatively small samples such as $n = 50$.

In contrasts, there was no impact of specifying aligned priors (compared to diffuse priors). Across the indices examined in Study 1, specifying informative aligned priors did not result in better model fit for the correctly specified model, nor did it increase the probability that the correctly specified model was selected when compared to several misspecified models. This lack of positive impact can be explained by the level of informativeness of the aligned prior. In particular, the prior standard deviation matched the precision of the data likelihood at $n = 50$ (see footnote 7), and in turn was less precise than the data likelihood at $n = 100$, 250, and 500. In addition, the prior agreed with the data likelihood, as it was centered on the population value. Under these conditions, the posterior samples will be similar to those based on diffuse priors. This finding is in line with the limited previous research that looked at using informative small-variance priors on cross-loadings (Liang, 2020) or informative priors for factor-loadings in a full SEM (Cain & Zhang, 2019).

These findings have important implications for applied researchers with access to limited sample sizes, who are often advised to specify informative priors for at least some of the parameters in their SEM (Depaoli, 2014; Lee, 2007; McNeish, 2016; Smid, Depaoli, et al., 2019; Smid & Winter, 2020). Some previous research has shown that informative priors generally result in less biased parameter estimates compared to diffuse priors, even if the informative priors diverge from the population values (e.g., Depaoli, 2014). Moreover, in the Bayesian framework, priors that diverge from the observed data are not inherently problematic. They simply reflect a disagreement between the prior knowledge about the parameter and the new evidence provided through the observed data. However, if a researcher is interested in model fit assessment or model selection, that trust in informative priors, even if they might disagree with the data, may not be merited. First, if priors are informative and perfectly aligned, they do not improve the researcher's ability to assess model fit or selection. Second, and more likely, if priors are informative, but they diverge from the unknown population parameter distribution, the model fit indices may indicate that fit is poor even when the correct model is specified. This finding points towards an important avenue for future research: examining methods for detecting disagreement between the priors and the data, so that applied researchers are aware and can adjust their reliance on model fit indices accordingly.

3.4.5 Conclusion

With the introduction of the Bayesian approximate fit indices and an improved version of the PPP-value to M*plus*, researchers have access to a host of new options for model fit and selection in Bayesian SEM. Based on the results of Study 1, researchers should be careful to rely on the approximate fit indices for model fit assessment of a single model. While the 90% credible intervals allow for the attractive option of inconclusive model fit, the inconclusiveness is not solely due to model misspecification, but may also be caused by the overall sample size, presence of missing data, or prior specification. Instead of focusing on a single model, researchers should compare multiple models to find a model that fits their data best. Furthermore, if a researcher uses informative priors in their model and finds that model fit is poor, they should examine whether this result may be related to any disagreement between their priors and the observed data.

Chapter 4

Study 2: Detecting Prior-Data Disagreement in Bayesian Structural Equation Modeling

4.1 Introduction

The choice of prior specification plays an important role in any Bayesian analysis. In Bayesian estimation, the information provided through the prior distribution is combined with information provided by the observed data likelihood to form the posterior distribution. Ideally, the prior distributions and the data likelihood support each other and tell the same story. But the prior and the data do not always agree. Prior-data disagreement occurs when the researcher's prior knowledge is not in agreement with the evidence provided by the data (Evans & Moshonov, 2006a).[10] Previous simulation studies on prior-data disagreement in structural equation model (SEM) estimation show that diverging priors introduce bias in the posterior distributions, especially when the priors are informative or the sample size is small (e.g., Depaoli, 2014; Dingjing Shi & Tong, 2017; Smid, Depaoli, et al., 2019). Thus, an essential next step in research on Bayesian SEM is to investigate methods to identify prior-data disagreement.

Several methods for identifying prior-data disagreement have been suggested in the literature. Examples include the Data Agreement Criterion (DAC; Bousquet, 2008), Bayes Factors (BFs; Kass & Raftery, 1995), and prior-predictive checks (PPC; Evans & Jang, 2010). As will become apparent, each method answers a slightly different question about the match between the prior and the data. Moreover, each of these approaches is easily applied to the SEM context. However, the question of their effectiveness in detecting prior-data disagreement in the SEM context has not been studied before. To address this gap in the literature, the second study of this dissertation focuses on the ability of these three methods to detect prior-data disagreement when estimating SEMs. For this purpose, I will focus on a model for which researchers are likely to incorporate informative priors: the latent growth model (LGM). This study aims to provide researchers who use SEM with concrete recommendations about which (if any) approach for detecting prior-data disagreement they should use.

The remainder of this section will be organized as follows. First, I will further introduce the concept of prior-data disagreement, after which I will present the three indices for detecting prior-data disagreement that are the focus of the current study. This

[10] It should be noted that prior-data disagreement would not pose a problem to a true subjective Bayesian, who would simply update their prior belief. However, prior-data disagreement can cause computational issues for the pragmatic evidence-based subjective Bayesian, which is why it is the focus of the current investigation.

subsection is followed by a discussion of the existing literature on the impact of divergent priors on SEM parameter estimation. Next, I will introduce and specify the LGM for which the three prior-data disagreement indices' performance will be examined through a simulation design.

4.1.1 Prior-Data Disagreement

Prior-data disagreement is generally discussed in the context of informative priors. However, it should be noted that the prior and data can interact in surprising ways even if the prior is considered diffuse (Smid & Winter, 2020; van Erp et al., 2018). Figure 40 provides a basic overview of different levels of prior-data (dis)agreement. In panel A, there is prior-data agreement: the prior (light grey) and the data likelihood (darker grey) overlap completely. In panel B, there is some prior-data disagreement. The distributions still overlap, but the center of the prior distribution is shifted to the left, indicating that the prior belief is centered around a lower value. Finally, in panel C, there is clear prior-data disagreement. The prior and data likelihood distributions have almost no visible overlap and cover completely different values on the *x*-axis.

It should be noted that the precisions of the prior distribution and the distribution created by the data likelihood do not need to be the same. Either source of information may be more or less precise. One can imagine a situation where the two distributions' centers align, but the prior distribution is narrower (conveying a more precise prior belief) or wider (conveying a less precise prior belief).

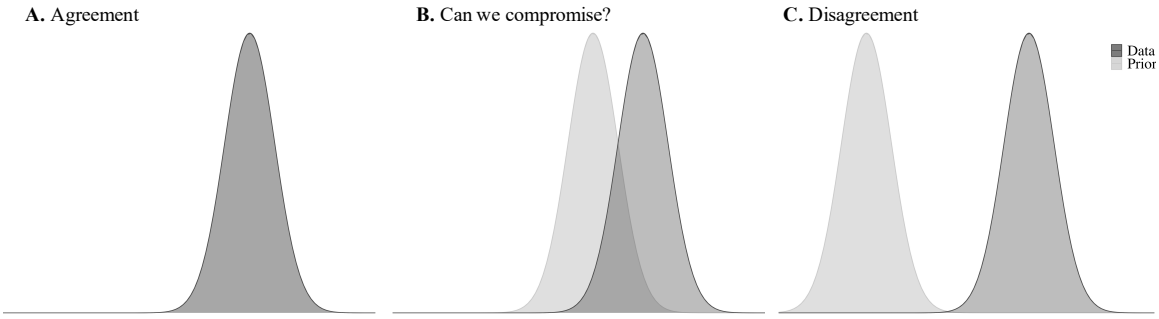


Figure 40. Illustration of different levels of prior-data (dis)agreement.

Concern about prior-data disagreement appears to have emerged in research looking at clinical trial study designs (Spiegelhalter et al., 1994; Young & Pettit, 1996). This field recognized the benefits of Bayesian estimation for building knowledge in an iterative manner but warned against the reliance on one prior specification. Instead, they recommended relying on a *community* of priors and examining prior-data disagreement for each specification (Kass & Greenhouse, 1989; Spiegelhalter et al., 1994). This process should be used as a diagnostic tool to indicate that two sources of information (the prior and the data) do not agree about a quantity of interest (Young & Pettit, 1996). Although it is easy to take evidence of disagreement as evidence that the prior must be wrong, there are many reasons for disagreement to emerge. For example, the problem may lie with the collected data. Perhaps the sample is not representative of the population about which the model was hypothesized (Veen et al., 2020). Alternatively, the problem

may reside with the research design. If the prior was based on a previous study that the current study is attempting to replicate, then disagreement between the prior and the data may indicate that the studies are different in some important ways (Young & Pettit, 1996). Finally, the problem may lie with the priors themselves. The sources upon which the priors were based, such as clinical experts, may not have been accurate (Spiegelhalter et al., 1994). Thus, if a researcher finds prior-data disagreement, they need to examine all aspects of their study closely.

As interest in detecting prior-data disagreement increased, researchers developed various indices to capture different expressions of prior-data disagreement. The next section will introduce three of these indices.

4.1.2 Indices for Quantifying Prior-Data Disagreement

This section will introduce three approaches to detecting prior-data disagreement: the DAC, BFs, and prior-predictive checking.

4.1.2.1 Data Agreement Criterion

The DAC represents the distance between a prior specified by the researcher and a diffuse reference prior (Bousquet, 2008). Before defining the DAC as an expression, it is important to introduce its main ingredient: the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951).[11] De KL divergence measures the loss of information if a reference distribution (π_1) is approximated by another distribution (π_2). The larger the discrepancy, or distance, between the two distributions, the greater the loss of information. The KL divergence is calculated as follows:

$$KL(\pi_1||\pi_2) = \int_{\Theta} \pi_1(\theta) log \frac{\pi_1(\theta)}{\pi_2(\theta)} d\theta, \tag{29}$$

where || stands for divergence, or the distance between the reference and the specified distributions, Θ is the parameter space for parameter θ, $\pi_1(\theta)$ is the reference distribution, and $\pi_2(\theta)$ is the distribution that attempts to approximate the reference distribution.

We can use the KL divergence to compute the DAC. In short, the DAC is the ratio of two KL divergences. Each KL divergence is based on the same reference distribution but uses a different distribution (i.e., a different prior) to approximate the reference distribution. To examine prior-data disagreement, the reference distribution is a posterior distribution $\pi^J(\theta|y)$ based on a benchmark prior $\pi^J(\theta)$ such that the data completely dominate the posterior distribution. For that reason, the benchmark prior is generally a diffuse or uninformative prior. To ensure that the KL divergence is well defined, this prior should be a proper prior. Figure 41 illustrates how the relationship between the reference distribution $\pi^J(\theta|y)$ and its associated KL divergence is different for the benchmark prior $\pi^J(\theta)$ (left panel) and a researcher-specified prior $\pi(\theta)$ (right panel). Two lower plots show the priors (dashed lines) and the reference posterior (solid line)

[11] Other distance measures can also be used, however the KL divergence appears to perform best (Lek & van de Schoot, 2019).

and the two upper plots show the KL divergence, which equals the shaded area under the curve. The DAC, using the reference posterior $\pi^J(\theta|y)$, the benchmark prior $\pi^J(\theta)$, the data y, and a chosen second prior $\pi(\theta)$, can be computed as follows:

$$DAC = \frac{KL[\pi^J(.|y)||\pi]}{KL[\pi^J(.|y)||\pi^J]}. \tag{30}$$

If the chosen prior conflicts less with the data than the benchmark prior, DAC < 1. If the chosen prior conflicts more with the data than the benchmark prior, DAC > 1. For the scenario depicted in Figure 41, the DAC is 1.15, slightly above the cutoff value. A larger DAC implies a more impactful prior-data conflict. The DAC will be > 1 for two reasons: (1) the researcher-specified prior is placed in a region of the parameter space that is far removed from the data (conflict in location), or (2) the researcher-specified prior is far more precise than the information from the data (conflict in information; Bousquet, 2008; Lek & van de Schoot, 2019). The specific question addressed with the DAC is: *Is this prior a good representation of the information present in the data about parameter θ?*

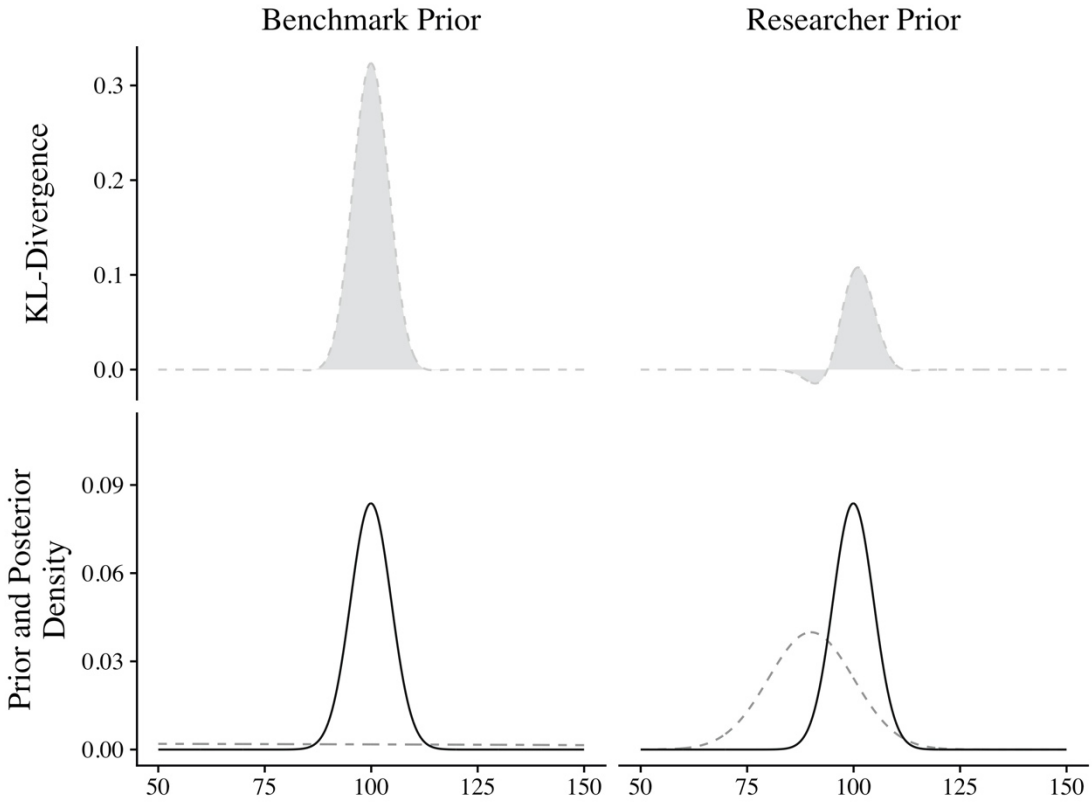


Figure 41. Illustration of the components that make up the DAC. Note the difference in *y*-axis scale for the upper and lower plots in the figure. Dashed line reflects the prior, solid line reflects the posterior.

For single-parameter models, the DAC can generally detect prior-data disagreement (Lek & van de Schoot, 2019; Veen et al., 2018). The DAC has also been used in an LGM context to compare the accuracy (compared to the observed data) of several prior distributions elicited from experts (Veen et al., 2020). In addition, it is relatively fast to compute the DAC for a large range of parameters and researcher-specified priors. This is because the computation of the DAC requires only one actual posterior distribution: the reference distribution, $\pi^J(\theta|y)$. All other (prior) distributions can simply be plugged into the formula. This makes the DAC an attractive option for SEMs, which often consist of many parameters. The main challenge of the DAC is that a benchmark prior needs to be selected for all parameters in the model. Although it is relatively straightforward to specify a diffuse prior for a single parameter, we know that diffuse priors in SEMs can interact in unexpected ways and produce a posterior distribution that does not accurately reflect the data (Depaoli, 2013; Depaoli & Clifton, 2015; Smid & Winter, 2020; van Erp et al., 2018).

4.1.2.2 Bayes Factors

A second way of looking at prior-data disagreement is through BFs. A BF can be computed by taking the ratio of the marginal likelihoods of two Bayesian analyses:

$$BF = \frac{m_1(y)}{m_2(y)}, \tag{31}$$

which provides the odds of some model M_1 versus model M_2 (Kass & Raftery, 1995). The marginal likelihood can be expressed as:

$$m(y) = \int_{\Theta} f\,(y|\theta)\pi(\theta)d(\theta), \tag{32}$$

for data y, parameter θ, and prior $\pi(\theta)$. As can be seen, the marginal likelihood is affected by the data, the model, and the prior (Kass & Raftery, 1995; C. C. Liu & Aitkin, 2008; Vanpaemel, 2010). Although the influence of the prior on the marginal likelihood is seen as a disadvantage by some (e.g., Gelman, 2008; Kass & Raftery, 1995), this sensitivity can be used to our advantage to detect prior-data disagreement. To compare different prior distributions, we must keep the data and the model constant across prior specifications. That way, the only difference in the marginal likelihood arises from a change in prior specification.

If we specify model M_1 using benchmark priors (i.e., $m^J(y)$), as with the DAC, and model M_2 using chosen priors (i.e., $m(y)$), then we can assess whether the marginal likelihood provides more support for the benchmark prior (BF > 1; prior-data disagreement) or the chosen prior (BF < 1; prior-data agreement). Evidence in favor of the benchmark or researcher-specified prior becomes stronger as the BF moves further away from 1 (Herbert Hoijtink et al., 2019). Thresholds for concluding "positive" or "strong" evidence have been suggested for the BF (e.g., 3 and 20 respectively; Kass & Raftery, 1995). However, the BF's sampling distribution depends on the model and prior specification, making universally applied cutoff values problematic (García-Donato &

Chen, 2005). The BF should be used simply to quantify the strength of support for one model compared to a second model (Herbert Hoijtink et al., 2019). Thus, the question addressed with the BF is: *Does this prior result in a marginal likelihood that is more supportive of the data than a benchmark prior?*[12] Alternatively, the BF can also be used to compare several pairs of chosen prior specifications to answer the question: *Does this prior result in a marginal likelihood that is more supportive of the data than some other prior?*

For a single parameter, the BF and DAC both assess one researcher-specified prior. However, for a model with multiple parameters, the BF assesses the entire collection of researcher-specified priors, whereas the DAC is computed for each parameter separately. The BF's focus on the overall model may be desirable, as we know that a prior for one parameter may affect the posterior of another parameter (Depaoli et al., 2020). The BF considers these interactions across priors and posteriors and provides an assessment of the entire model. A drawback of this approach is that it may be challenging to identify which prior(s) drive the prior-data disagreement. In addition, the BF approach is computationally more demanding, as the model needs to be estimated for each prior specification. Moreover, similar to the DAC, a series of diffuse benchmark priors needs to be specified for the reference model.

In addition to these characteristics of the BF, the researcher should also consider an important distinction between BFs and DACs: BFs are less likely than DACs to flag researcher-specified priors that are too precise as prior-data disagreement (conflict of information; Veen et al., 2018). The reason behind this becomes evident if we express the DAC in terms of the BF (Bousquet, 2008; Veen et al., 2018):

$$\text{DAC} = \frac{m^J(y)}{m(y)} exp\left\{KL[\pi^J(\cdot\,|y)||\pi(\cdot\,|y)]\right\} = \text{BF}\exp\left\{KL[\pi^J(\cdot\,|y)||\pi(\cdot\,|y)]\right\}. \quad (33)$$

Equation (33) shows us that the DAC has an additional penalty term that multiplies the BF by the KL divergence between the reference posterior and the posterior based on the researcher-specified prior. Note that this expression of the DAC uses the posterior based on the researcher-specified prior, $\pi(\cdot\,|y)$, instead of simply the researcher-specified prior, π, as in the original expression of the DAC in the previous section. It may depend on the situation whether the DAC or the BF should be preferred for detecting prior-data disagreement. The DAC and BF are likely to disagree for analyses based on small samples that include informative priors. This disagreement emerges because it is likely that the prior conveys more information about the parameter than the observed data. According to the DAC, this makes the prior overly precise, which will likely indicate prior-data disagreement. In contrast, the BF may indicate prior-data agreement, perhaps even supporting the researcher-specified priors over the benchmark priors. In this example, only the DAC makes the researcher aware that their prior specification may be the driving force behind the posterior distribution. Thus, it may be that researchers need

[12] As the marginal likelihood is likely support a diffuse prior over any informative prior, it is likely that several prior specifications need to be compared to the diffuse benchmark prior to gain insight into the relative ranking of the researcher-specified priors.

to assess multiple data-disagreement indices to become aware of different forms of disagreement (e.g., disagreement in precision or location).

4.1.2.3 Prior-Predictive Checking

A third way of investigating prior-data disagreement is through prior-predictive checks (Box, 1980; Evans & Moshonov, 2006a; Gelman et al., 2017). Prior-predictive checks are based on the prior predictive distribution, which is generated by evaluating the marginal likelihood without including the observed data. In other words, we run a Bayesian analysis for which the iterative output for parameter θ is only affected by the prior distribution π. We can then use the prior predictive distribution to generate random samples of y, often denoted as y^{sim}. This is done by first simulating parameters according to the priors

$$\theta^{sim} \sim \pi(\theta), \tag{34}$$

and then simulating data according to the sampling distribution given the simulated parameters

$$y^{sim} \sim \pi\left(y \middle| \theta^{sim}\right). \tag{35}$$

These random samples form a distribution of all possible samples that could occur if the model and prior specification are true (van de Schoot et al., 2021). If the prior aligns with the data, then the random samples y^{sim} will form a prior predictive distribution that is similar to the true data-generating distribution. To assess whether prior-data disagreement exists, these random samples can be used to create a fixed distribution P_T of some relevant piece of information about the observed variable y (e.g., its mean) called $T(y_0)$.[13] The fixed distribution P_T is then used to examine whether $T(y_0)$ is surprising (Evans & Moshonov, 2006a). A *prior* predictive *p*-value (Evans & Jang, 2010; Evans & Moshonov, 2006a) can quantify the unexpectedness of $T(y_0)$, by comparing the value of the density p_T of T at $T(y_0)$ with other possible values:

$$P_T = P\left(p_T(t) \leq p_T\left(T(y_0)\right)\right). \tag{36}$$

Ideally, the prior predictive *p*-value should be close to 0.5, which indicates that about half the density p_T is above (and below) $T(y_0)$. The specific question that prior-

[13] This piece of information should be a minimally sufficient statistic. A statistic T is sufficient if knowing the value of T leads to an estimate of θ that is just as accurate as an estimate based on the entire random sample that T is based on. Examples include the sample mean for a normally distributed θ with a known variance, or the maximum observed value for a θ that represents the upper bound of a uniform distribution. For a normal distribution with an unknown mean μ and variance σ^2, the minimal sufficient statistic is $T(\bar{y}, s^2)$ (i.e., sample mean and sample variance estimate). Since the distribution of s^2 does not depend on μ, it is also possible to assess each statistic separately through the marginal prior predictive distribution. By looking at the mean and variance separately, it is possible to differentiate between prior-data conflict that arises from the location of the data and the spread of the data (Evans & Moshonov, 2006a).

predictive checks address is: *Is this prior able to predict my observed data well?* Or: *Are my observed data unexpected under this prior specification?*

Using prior-predictive checks such as the prior predictive *p*-value has several advantages. The approach is conceptually straightforward and does not require the specification of benchmark priors or a reference posterior. The method is more computationally intensive than the DAC, but less intensive than the BF approach. The main challenge lies in the specification of the minimally sufficient statistic T that provides relevant information to assess how unexpected the observed data are under the selected prior specification. Moreover, Young and Pettit (1996) argued that measures such as the prior-predictive *p*-value do not differentiate between two priors centered over the true parameter space that only differ in terms of precision. This indifference distinguishes the prior-predictive *p*-value from the DAC (which may prefer the less precise prior if it is more in line with the precision provided through the likelihood) and the BF (which prefers the more precise prior).

4.1.3 The Impact of Prior-Data Conflict on SEM Estimates

Being aware of prior-data conflict is valuable in its own right (Evans & Moshonov, 2006a; Young & Pettit, 1996). However, this sense of awareness might quickly be followed by a new question: *Does it matter?* The disagreement between the prior and the data may or may not affect inferences based on the posterior distribution. Particularly for larger samples, the influence of the prior may be minimal (Evans & Moshonov, 2006a; Liang et al., 2020). In this section, I will review the literature on the influence of priors on SEM estimates to examine if detecting prior-data disagreement is more important in some situations than in others.

Numerous studies have looked at the effect of diverging prior distributions on the accuracy and efficiency of posterior parameter estimates (e.g., Depaoli, 2014; Depaoli et al., 2017; Finch & Miller, 2019; Holtmann et al., 2016; Marcoulides, 2018; Miočević et al., 2020; Dingjing Shi & Tong, 2017; Smid, Depaoli, et al., 2019). Several patterns have emerged from this research. First, the impact of diverging priors may be limited if these diverging priors are not *too* precise (Depaoli, 2014; Finch & Miller, 2019; Holtmann et al., 2016; Miočević et al., 2020). In fact, across these studies, divergent weakly informative priors often resulted in more accurate and efficient posterior parameter estimates than diffuse priors. In contrast, highly informative, divergent priors tended to result in severely biased posterior estimates (Depaoli, 2014; Depaoli et al., 2017; Marcoulides, 2018; Dingjing Shi & Tong, 2017; Smid, Depaoli, et al., 2019), in some cases also causing bias in other model parameters (Depaoli, 2014; Holtmann et al., 2016). Moreover, these informative divergent priors resulted in biased posterior estimates that were highly efficient (i.e., had narrow 95% credible intervals), which increases the risk of drawing inappropriate inferences regarding the presence or absence of an effect (Dingjing Shi & Tong, 2017).

Another worrisome finding across multiple studies is that access to sample sizes that most would consider large enough to overwhelm the prior did not wholly diminish the negative impact of informative divergent priors (Depaoli, 2014; Depaoli et al., 2017; Holtmann et al., 2016; Dingjing Shi & Tong, 2017). For example, Depaoli (2014) and Depaoli and colleagues (2017) showed that the problematic impact of informative

divergent priors on growth mixture model parameter estimates was still present for an overall sample size of n = 600 or 800. Similarly, Holtmann and colleagues (2016) examined the impact of informative divergent priors on multilevel SEM parameter estimates. They found that bias in the posterior estimates persisted even for larger samples with 200 clusters of 6 observations (total overall n = 1200). Finally, Shi & Tong (2017) found that severely divergent priors resulted in biased posterior estimates in a latent basis growth model across all included sample size levels (n = 50 to 500). Thus, it may be imperative to assess prior-data disagreement if highly informative priors are specified. Depending on their accuracy, these priors can result in the biggest gains or losses in accuracy and efficiency, even with larger samples.

There is one additional group of priors that needs to be mentioned here: diffuse priors. As briefly mentioned above, diffuse priors may be more problematic than weakly informative divergent priors in SEM estimation (Baldwin & Fellingham, 2012; Depaoli, 2014; Depaoli et al., 2017; Depaoli & Clifton, 2015; Finch & Miller, 2019; Holtmann et al., 2016; Smid & Winter, 2020; van Erp et al., 2018). Diffuse priors can become unexpectedly influential for analyses based on small samples, where the observed data contribute a limited amount of information. While diffuse priors are not typically thought of as presenting prior-data disagreement, it is important to keep their potential impact on the posterior estimates in mind. Thus, a lack of prior-data disagreement does not guarantee that a prior does not have other inadvertent effects (Smid & Winter, 2020). Out of the three indices for detecting prior-data disagreement included in the current study, the BF may be the only index that will show a preference for an informative prior over a diffuse prior.

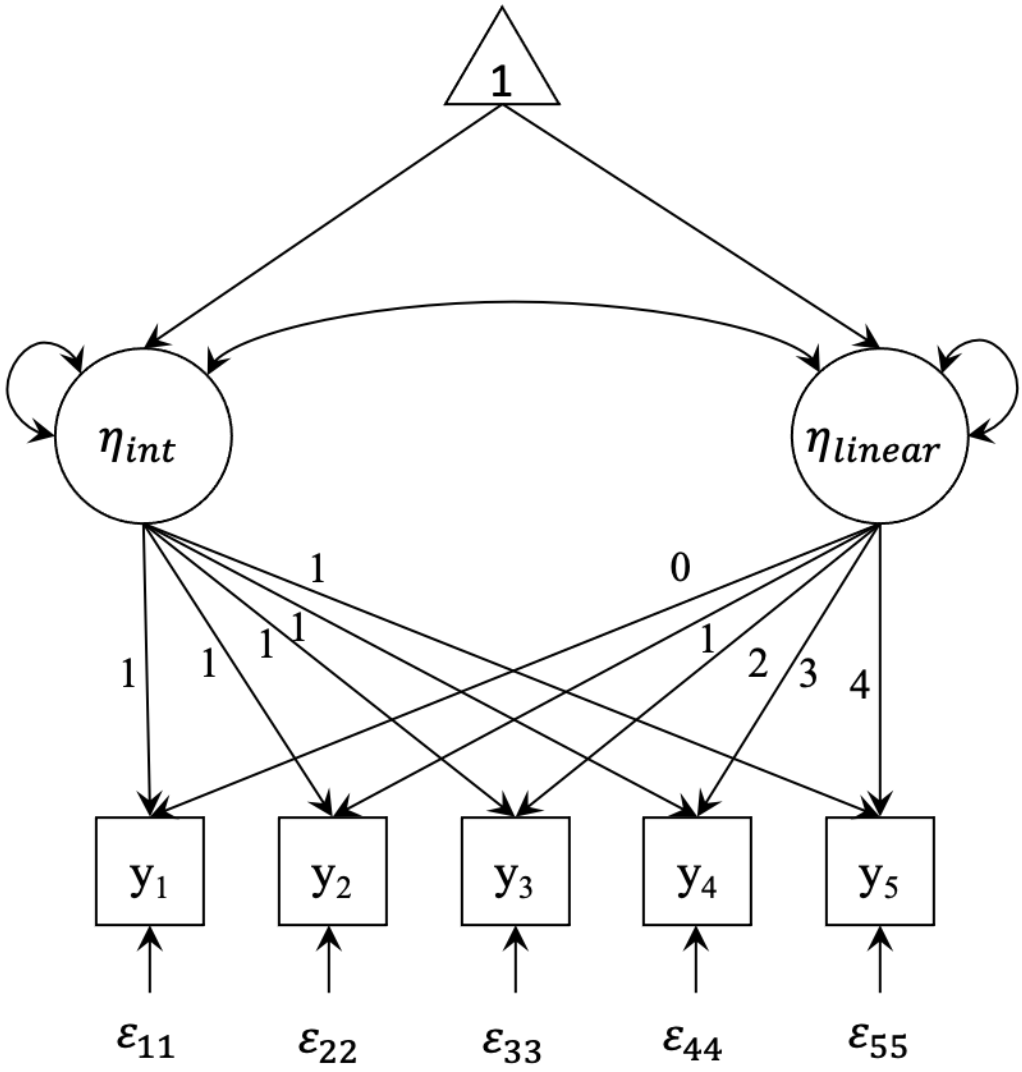
4.1.4 Model Examined in the Current Study

To examine if the DAC, BF, and prior-predictive p-value can detect prior-data disagreement in an SEM, the current study will focus on a commonly used SEM: the LGM. Specifically, the population model is an LGM with 5 time points (Figure 42). The model can be expressed in the following matrix form:

$$\mathbf{Y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}, \tag{37}$$

where $\mathbf{Y}$ represents a vector of repeated measures variables, $\boldsymbol{\eta}$ represents a vector of latent variables (the growth parameters), and $\boldsymbol{\Lambda}$ is a fixed loading matrix relating the growth parameters to the observed outcomes. The first column is related to the intercept and is a column of ones. Each additional column represents a specific slope (e.g., linear, quadratic). For the population model in Figure 41, these values are 0, 1, 2, 3, 4 for the linear slope. This means that the intercept is located at the first time point and timepoints are equally spaced (but this spacing can be altered if desired). Finally, $\boldsymbol{\varepsilon}$ represents a vector of residuals. Further, $E(\boldsymbol{\eta}) = \boldsymbol{\alpha}$, a vector of means of the latent variables, $\boldsymbol{\Phi}$ is the covariance matrix of the latent variables (between-individual covariance matrix), $\boldsymbol{\Sigma}$ is the population covariance matrix, and $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ is the covariance matrix of residuals. This matrix is diagonal in the absence of residual covariances. Furthermore, we assume $Cov(\boldsymbol{\eta}, \boldsymbol{\delta}) = 0$. Following this, the covariance matrix of the observed data can be expressed as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\Phi}\boldsymbol{\Lambda}_{\boldsymbol{y}}' + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}. \quad (38)$$



$$\Phi = \begin{bmatrix} 1.00 & \\ 0.11 & 0.20 \end{bmatrix}$$

$$\alpha = \begin{bmatrix} 1.00 \\ 0.80 \end{bmatrix}$$

$$\Theta_{\varepsilon}\ (diagonal) = [1.00 \quad 1.42 \quad 2.24 \quad 3.46 \quad 5.09]$$

Figure 42. Path diagram and population parameters for Latent Growth Model (LGM).

4.2 Design

4.2.1 Population Values

Population values for the LGM are included in Figure 41 and are based on Bauer and Curran (Bauer & Curran, 2003). These values represent a scenario in which the latent growth factors (i.e., intercept and linear slope) account for 50% of the variance in each of the observed variables, which was deemed appropriate for the current investigation.

4.2.2 Sample Size

The impact of a selected prior distribution changes as a function of the sample size. On the one hand, priors that may seem diffuse or noninformative can become highly informative when the sample size is relatively small (Mcneish, 2016). On the other hand, highly informative priors may have no impact on the posterior distribution when the sample size is relatively large (B. O. Muthén & Asparouhov, 2012). The current study examined the impact of sample size by including four different sample sizes: $n = 50$, $n = 100$, $n = 250$, and $n = 500$.

4.2.3 Prior Specification

To investigate the impact of prior specification I included 7 prior specifications for the intercept mean (Figure 43, panel A) and the slope mean (Figure 43, panel B). All priors followed a normal distribution, $N(\mu, \sigma)$, with mean hyperparameter μ and standard deviation hyperparameter σ. The first prior specification served as the benchmark prior for the DAC and BF computation, $N(0, 55)$.[14]

The next three prior specifications were all centered around the population values for the intercept mean (1.0) and slope mean (0.8), and they are labeled as "aligned". In Figure 42A and B, these priors are centered over the triangle (population value) and drawn with bold lines. These priors vary in the specification of the standard deviation (i.e., the precision of the prior). Starting with the middle level of precision (moderately informative), I selected a standard deviation of 0.2 for the intercept mean and 0.1 for the slope mean.[15] To vary the informativeness of the priors, I either multiplied the standard deviation of the moderately informative prior by 2 (weakly informative) or by 0.5 (strongly informative).

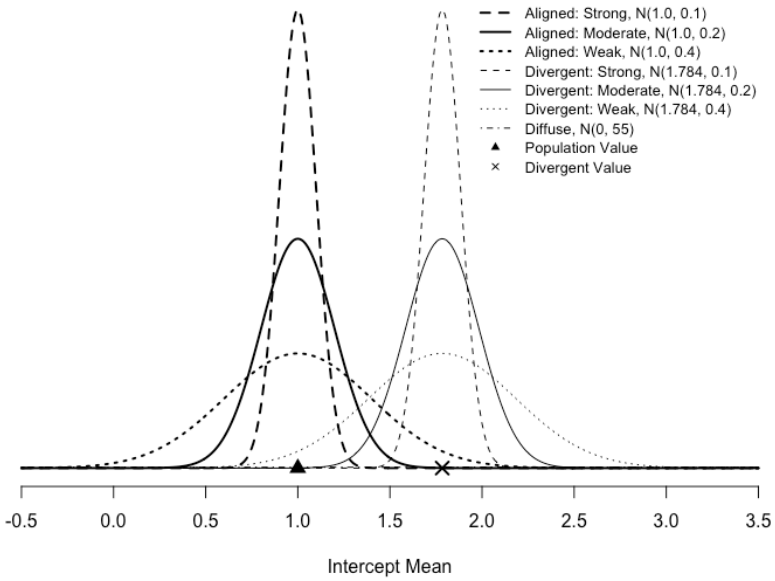
---

[14] This prior was selected after a pilot study demonstrated that this prior had minimal impact on the posterior distribution but did not result in convergence issues. Other priors investigated were: $N(0, 10)$, $N(0, 100)$, and $U(-100, 100)$.

[15] These values represent data dependent priors (DDP; Mcneish, 2016), which I selected by first generating 100 samples of $n = 50$ from the population model. Next, I used these 100 samples to estimate the population model with maximum likelihood estimation (MLE). I averaged the standard error estimates of the intercept mean and slope mean across the 100 samples. I used those values to specify the standard deviation hyperparameters of the moderately informative priors. I used the smallest sample size conditions of the main simulation design to find the standard error estimate that represented the highest level of uncertainty. The same values were used across all sample size levels in the main simulation. In applied research, DDPs are somewhat controversial, as the researcher technically double-dips by using their data to specify the priors that are subsequently used to analyze their data. However, they can aide in model estimation under certain circumstances (e.g., Mcneish, 2016).

The final three prior specifications were labeled as "divergent" and represented priors that diverged from the population value. In Figure 19A and B, these priors are

A. Prior Specifications for the intercept mean



B. Prior Specifications for the slope mean

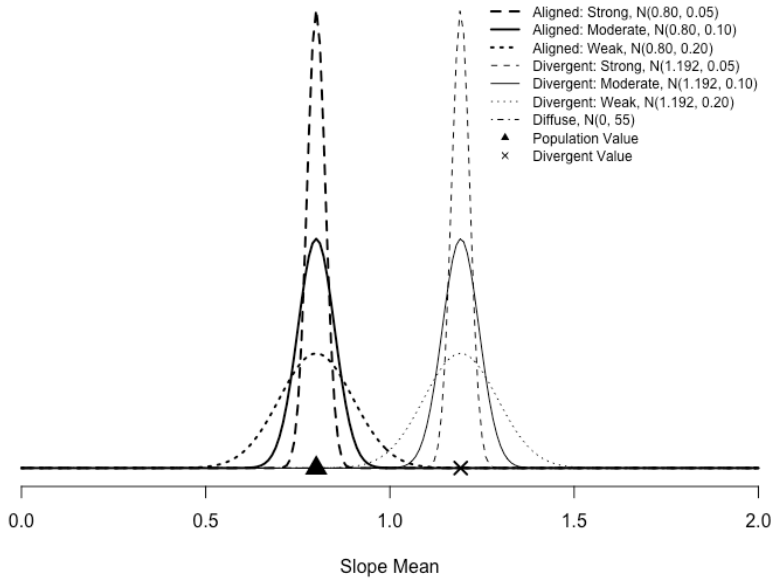


Figure 43. Prior conditions for the intercept mean (panel A) and slope mean (panel B).

centered over the cross and drawn with normal weight lines. The mean hyperparameter for these three priors was selected by finding the μ value for which a normal distribution had 5% overlap with the correct moderately informative prior. The standard deviation of the correct moderately informative prior was used in this process. This procedure resulted in the mean hyperparameter of the divergent priors being 1.784 and 1.192 for the intercept and slope mean respectively. The standard deviations of the divergent priors matched those of the correct priors. The 7 priors for each parameter are fully crossed, resulting in 49 prior conditions.

4.2.4 Data Generation

I chose the number of replications included in each cell of the simulation after assessing at what point the simulation converged. To ensure convergence of the simulation to a stable estimate, I examined cumulative average plots for all conditions. Based on these plots, 1,000 replications were sufficient to ensure that the simulation converged to a stable estimate across all simulation cells. I generated all data in R (R Core Team, 2019) using the package 'lavaan' (Rosseel, 2012). I simulated data separately for each of the four sample size conditions.

4.2.5 Bayesian Estimation

The R package 'rstan' (Stan Development Team, 2020) with its default sampler, NUTS (Betancourt, 2018; Hoffman & Gelman, 2014), was used for Bayesian estimation. Each model was estimated using four chains, the default number of chains used in the 'rstan' package. Each chain consisted of 10,000 iterations, with the first 5,000 iterations discarded as burn-in. I selected that number of iterations after testing several chain lengths for a select number of replications in each cell and inspecting the trace plots, the $\hat{R}$ convergence diagnostic, and ensuring that each parameter's ESS was > 1000 (Zitzmann & Hecht, 2019). To further ensure that convergence was obtained across all replications, I checked the $\hat{R}$ convergence diagnostic and effective sample size for all replications across all conditions.

For each prior condition, I estimated two models. The first model was a full Bayesian model that used the priors and the data to sample from the posterior distribution. The second model was a Bayesian model that did not include the observed data. This model was used to generate the prior-predictive samples that were used to compute the prior-predictive *p*-value. I executed all analyses through R. Output generated as a result of the simulations was also processed through R.

4.2.6 Outcomes of Interest

I extracted several types of results. First, I examined the relative bias and root mean-square error (RMSE) of the posterior estimates of the mean intercept and slope parameter to see how the prior specification affects the posterior estimates. In the current investigation, I used the rule of thumb that relative bias that exceeds |.10| reflects meaningful bias (Flora & Curran, 2004). It should be noted that this rule of thumb should not be blindly applied to any situation. A more meaningful definition of bias should be used if it is available. Further, for an unbiased parameter estimate, the RMSE reflects the sampling standard deviation of that parameter. For a biased parameter estimate, the

RMSE combines bias and variability into an overall measure of average error in the estimate. When comparing multiple estimates of the same parameter, estimates with lower RMSEs are preferred to estimates with larger RMSEs.

Second, I assessed the performance of the three indices for detecting prior-data disagreement in various ways. Specifically, for the DAC and the BF, I examined the average value of the index, the proportion of replications for which a certain prior resulted in prior-data disagreement, and the number of times each prior specification was selected as the "best" prior. The DAC has a clear cutoff value denoting prior-data disagreement (DAC > 1). However, for the BF, the question may be asked at what point its value reflects strong enough evidence of prior-data disagreement. Rules of thumb have been suggested in the past, such as BF > 3 or > 10 (Jeffreys, 1961; Kass & Raftery, 1995). However, research has shown that the BF depends on, for example, the sample size (Morey & Rouder, 2011). Thus, in the current study, I performed a cut-point analysis on a random sample consisting of 50% of the replications in each sample size condition to find the optimal cutoff value for each sample size. As cut-point analyses differentiate between two outcomes, the prior specifications were classified such that specifications with at least one divergent prior represented disagreement. To find the optimal cutoff value, I used the R package 'cutpointr' (Thiele, 2021) and used a method that optimized the sum of specificity and sensitivity. I used bootstrapping (with 200 draws) to find a robust BF cutoff value for each sample size. Next, I used the cutoff values to assess prior-data disagreement in the remaining 50% of the replications. The replications were split into two parts to prevent overly accurate results by using the same values twice.

For the prior-predictive *p*-value, I used the mean of each observed variable as the statistic T. A prior-predictive proportion close to .5 implied that the model and priors could recover the observed variable means well, whereas a prior-predictive proportion < .05 or > .95 implied that the model and priors were not able to recover the observed variable means well. To assess the performance of this index, I examined the average value of the index, the proportion of replications for which the prior-predictive *p*-value indicated prior-data disagreement, and the number of times each prior specification resulted in a prior-predictive *p*-value that was closest to the ideal (i.e., .5).

4.3 Results

The results are organized as follows: I first discuss findings regarding convergence of each of the replications in the simulation. This is followed by an assessment of the relative bias and RMSE of the posterior estimates to determine when prior-data disagreement affects the posterior estimates. Next, I examine the results for each prior-data disagreement index sequentially.

4.3.1 Assessing Convergence

For each replication in the simulation, I extracted the largest $\hat{R}$ and smallest ESS to assess convergence and precision of the posterior distributions sampled with NUTS. For n = 50, 28 replications resulted in $\hat{R}$ > 1.05 for at least one of the 49 estimated models. For n = 100, 3 replications resulted in $\hat{R}$ > 1.05 for at least one of the 49 estimated models. For n = 250 and 500, zero replications resulted in $\hat{R}$ > 1.05.

For n = 50, models with at least one diverging prior were more likely to result in non-convergence. For n = 100, the prior specification did not appear related to whether convergence was reached. Furthermore, after inspecting the ESS across sample sizes and prior specifications, the lowest ESS was always associated with one of the residual variance parameters. Replications for which non-convergence was found (based on $\hat{R}$ > 1.05) for at least one model were completely excluded from further examination of the results. Cumulative average plots were inspected after removing these replications, to ensure that the simulation was still converged for each sample size condition.

4.3.2 When Should the Prior-Data Disagreement Indices Flag Disagreement?

The impact of the prior changes as a function of its level of informativeness (e.g., how narrow the prior is) and the sample size of the observed data. Thus, detecting prior-data disagreement may be especially important under conditions in which the prior is likely to affect the posterior estimates. For that reason, I first assessed under what conditions these indices *should* flag prior-data disagreement. Here, I focused on the relative bias and RMSE of the two parameters for which the priors were altered (i.e., the mean intercept and mean slope), as the impact of their prior specification on their own posteriors is likely to be the largest (compared to their impact on other model parameters).

4.3.2.1 Relative Bias

Relative bias in the mean intercept and mean slope parameter estimates is depicted in Figure 44. Within the figure, the bias of the mean intercept is shown in the left column and the mean slope in the right column. The rows represent the sample size levels. Within each plot, relative bias is shown on the *y*-axis (with dashed lines representing |10%| bias), and each group of bars reflects a prior specification of the slope parameter. Within each group of bars, each bar reflects a prior specification of the intercept parameter. Both between and within groups of bars, the priors are ordered from optimal (i.e., aligned and strongly informative) to neutral (i.e., diffuse), to worst (i.e., divergent and strongly informative). The expectation is that parameter recovery worsens as we move from left to right within each group of bars and each plot.

Overall, relative bias decreased as the sample size increased. Focusing on the mean intercept estimate bias (left column), the results showed that if the intercept prior was divergent and moderately or strongly informative, the intercept estimate was positively biased (in line with the direction of the divergent intercept prior; lightest two bars). However, if the intercept prior was diffuse or aligned and weakly informative, and paired with the divergent strongly informative slope prior (right-most group of bars), the estimate was meaningfully biased in the negative direction for n = 50 and 100. This may reflect that the mean intercept estimate was forced down to balance the positively biased mean slope estimate (see right-most group of bars in right column). When n = 250 or 500, the mean intercept estimate was biased only with divergent strongly informative priors on both parameters.

Moving to the mean slope estimate bias (right column), we can see that the slope estimate reached a negative bias greater than -.10 with n = 50, but only if the slope prior was diffuse (center group of bars) and the intercept prior was divergent and strongly

informative (lightest bar). The mean slope estimate became positively biased when n = 50 or 100, with divergent moderately or strongly informative priors (left-most groups of bars), regardless of the prior specification of the intercept. When n = 250 or 500, the mean slope estimate was biased only with divergent strongly informative priors.

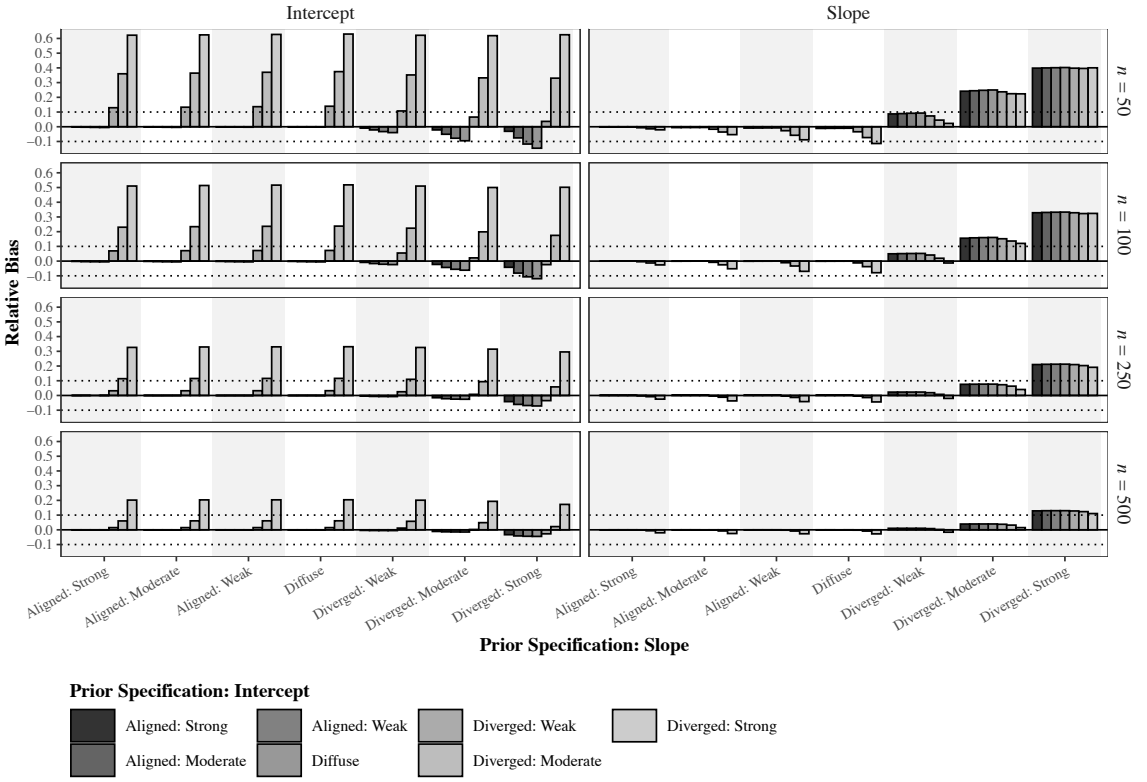


Figure 44. Relative parameter bias of the mean intercept and slope parameter across simulation conditions.

Based on these results, the prior-data disagreement indices need to consistently detect prior-data disagreement if divergent moderately or strongly informative priors are placed on both parameters *or,* for smaller sample sizes (i.e., n = 50 or 100), if a diffuse prior is placed on one parameter and a divergent strongly informative prior is placed on the other parameter. Moreover, specifying aligned strongly informative priors does not negatively impact the relative bias. In fact, an aligned and strongly informative prior resulted in the least biased posterior estimates for the parameter on which the prior was placed. Thus, the prior-data disagreement indices should not flag aligned and strongly informative priors as problematic.

4.3.2.2 RMSE

RMSE of the posterior estimates of the mean intercept and mean slope parameter estimates is depicted in Figure 45. This figure is organized in the same manner as Figure 44. Overall, RMSE values decreased as the sample size increased, following a similar pattern to the relative bias.

Focusing on the RMSE of the mean intercept estimate, the pattern and magnitude of RMSEs were nearly identical across different prior specifications for the slope (i.e., across groups of bars within each plot). Across the intercept prior conditions, the RMSE was lowest with aligned and strongly informative priors (left-most bar within each group). Further, the RMSE steadily increased with diffuse and divergent weakly informative priors. Between divergent weakly and moderately informative priors, the RMSE increased steeply and was highest for the divergent strongly informative prior.

Moving to the RMSE of the mean slope estimate, we can see that the pattern of results was independent of the mean intercept prior specification, as the bars within each group tended to be of similar magnitude. However, there did appear to be a small change for models estimated with a divergent and strongly informative mean intercept prior (the right-most bar within each group). This divergent prior inflated the RMSE if the slope priors were aligned or diffuse, while it reduced the RMSE if the slope priors were divergent. However, this effect was small compared to the effect of the mean slope prior specification itself. Across mean slope prior conditions, the RMSE increased slightly, moving from aligned and strongly informative to divergent and weakly informative. However, the RMSE increased more visibly if divergent and moderately or strongly informative priors were specified.

Thus, in terms of minimizing the RMSE, the prior-data disagreement indices need to consistently detect disagreement if priors placed on either of the parameters were divergent and moderately or strongly informative.

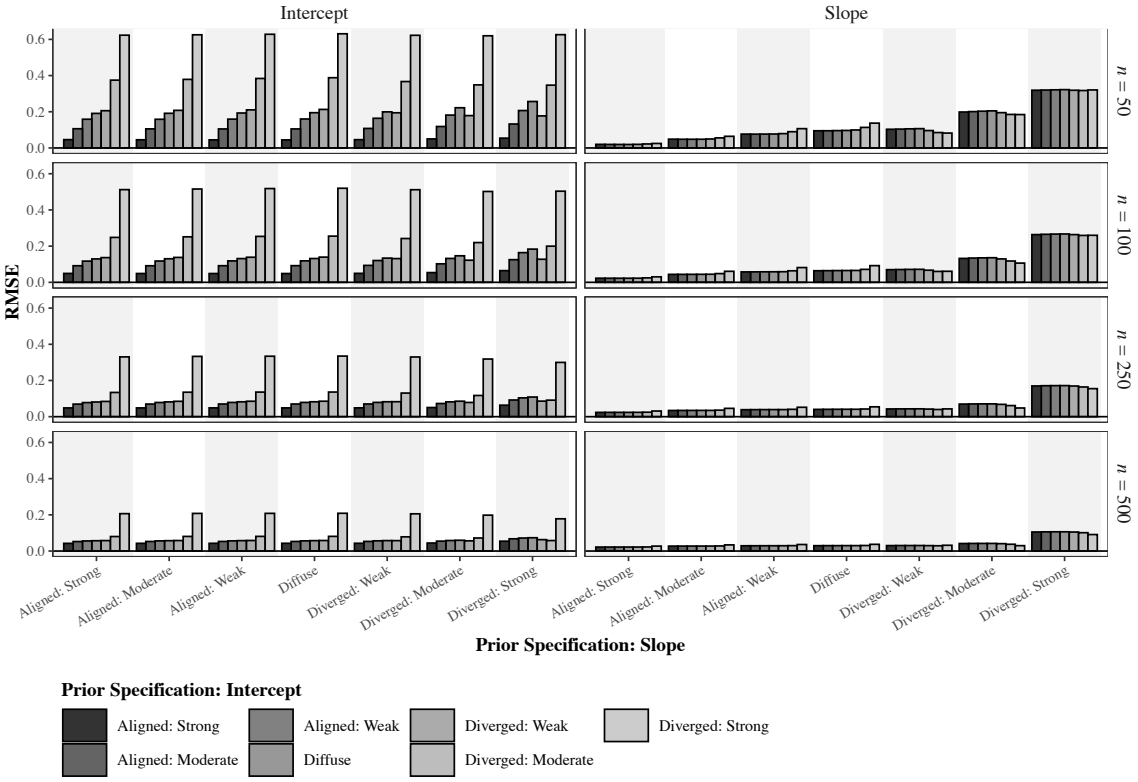


Figure 45. RMSE of the mean intercept and slope parameter across simulation conditions.

4.3.3 The DAC

4.3.3.1 Proportion of replications for which DAC > 1

Figures 46 and 47 show the proportion of DACs > 1 (indicating prior-data disagreement) for the intercept and slope prior, respectively. As the DAC is computed relative to a benchmark posterior based on an analysis with diffuse priors placed on all parameters, the DAC can be examined for each parameter independently. Proportions are grouped by prior specification and displayed for each sample size level. The specific proportion values are reported at the top of each plot for clarity.

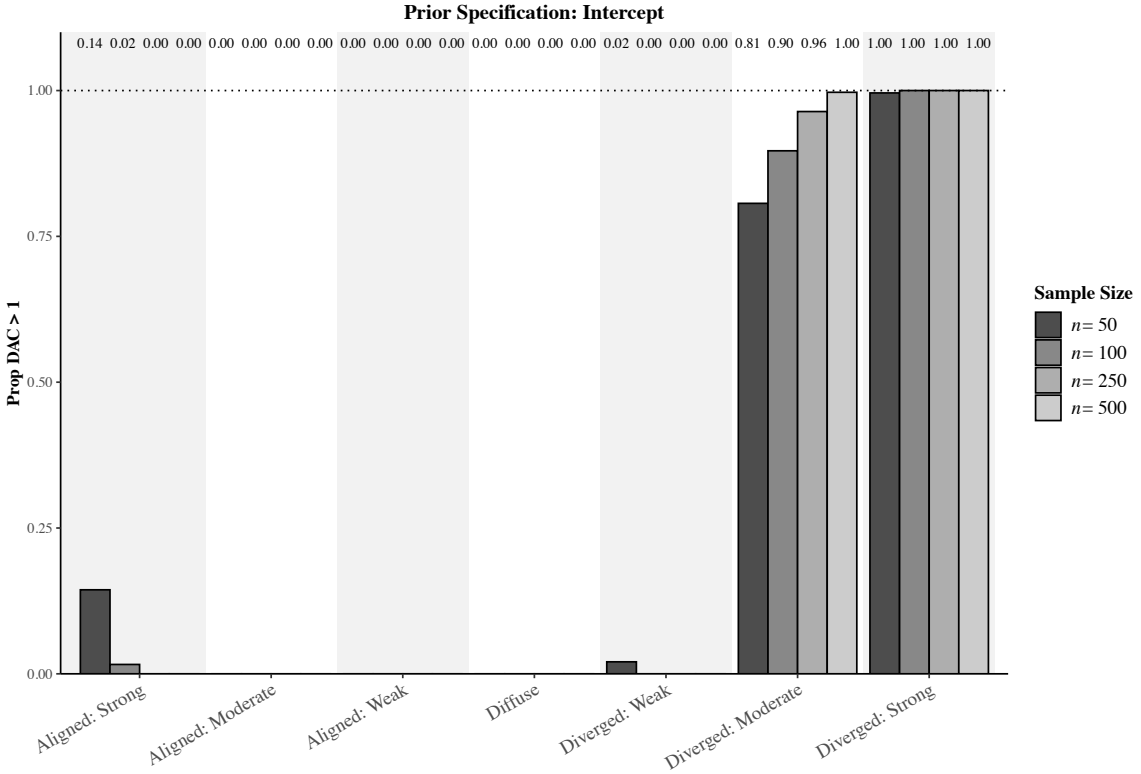


Figure 46. Proportion of DACs > 1 for the mean intercept prior across simulation conditions.

In these figures, we can see a similar pattern for both parameters. As expected, the DAC was never > 1 for diffuse and aligned weakly or moderately informative priors. For the aligned strongly informative priors, the DAC was > 1 for some replications, particularly for the smaller sample size condition. This result was expected, as the DAC also penalizes priors that are *too* informative relative to the information provided through the data. This effect disappeared as the sample size increased.

The DAC was unlikely to be > 1 for the divergent weakly informative prior condition. Moving from weakly to moderately informative divergent priors, there was a large increase in the proportion of DAC > 1, with over 75% of the replications resulting in DAC > 1 for both parameters. Almost all (or all for n = 100, 250, and 500) replications

using the divergent strongly informative priors resulted in a DAC > 1. Comparing Figure 45 to Figure 46, we can see that the DAC appeared slightly more sensitive to misspecification in the mean intercept prior compared to the mean slope prior.

Overall, the DAC could detect divergent priors that are strongly informative even for data with small sample sizes. The DAC appeared to become more sensitive to prior-data disagreement as the sample size increased. However, for small sample sizes, the DAC sometimes indicated prior-data disagreement if the prior was too informative for the data.

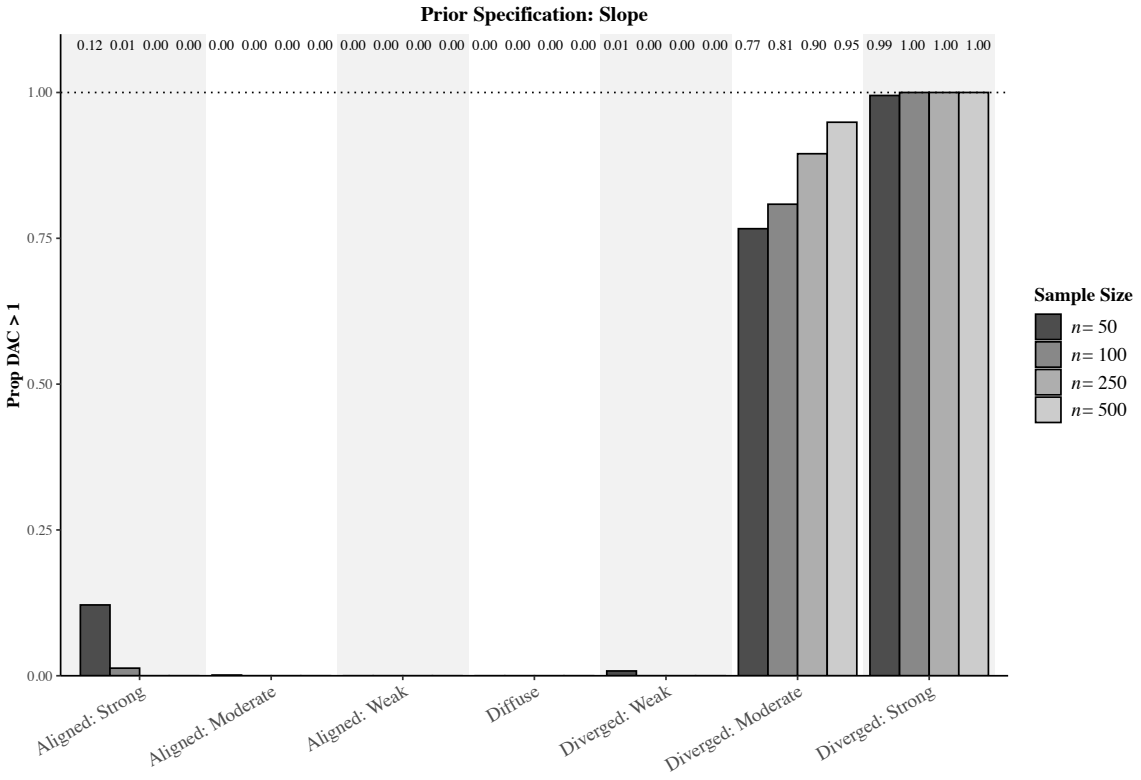


Figure 47. Proportion of DACs > 1 for the mean slope prior across simulation conditions.

4.3.3.2 Optimal Prior as Identified by the DAC

According to the DAC, the distribution with the smallest DAC value reflects a prior that is most in agreement with the data. Table 4 below presents the proportion of times each prior specification resulted in the lowest DAC for the mean intercept and slope parameters. As the sample size increased, the lowest DAC more often occurred for increasingly informative aligned priors. For n = 50 and 100, aligned and moderately informative priors were most often associated with the lowest DAC, whereas for n = 250 and 500, aligned and strongly informative priors were most often associated with the lowest DAC. The divergent priors rarely emerged as the optimal prior specification. Thus, if the DAC is compared across a set of prior specifications, it is highly unlikely that a severely divergent prior will emerge as the optimal specification.

Table 4. Proportion of times each prior specification was associated with the lowest DAC across simulation conditions.

Parameter	Sample Size		Aligned			Divergent		
		Diffuse	Weak	Moderate	Strong	Weak	Moderate	Strong
Intercept	50	0.000	0.284	0.693	0.000	0.022	0.001	0.000
	100	0.000	0.066	0.749	0.183	0.002	0.000	0.000
	250	0.000	0.002	0.189	0.809	0.000	0.000	0.000
	500	0.000	0.000	0.033	0.967	0.000	0.000	0.000
Slope	50	0.000	0.356	0.634	0.000	0.013	0.000	0.000
	100	0.000	0.068	0.854	0.078	0.000	0.000	0.000
	250	0.000	0.001	0.185	0.814	0.000	0.000	0.000
	500	0.000	0.000	0.043	0.957	0.000	0.000	0.000

4.3.4 The BF

For the BF, sets of prior specifications can be compared to a model with diffuse priors. Thus, each unique combination of intercept and slope prior specification could be examined.

4.3.4.1 Changes in the BF Values across Prior Specifications

For the BF, a larger value indicates more apparent prior-data disagreement. Table 5 presents the median BF values are reported in Table 6. I decided to focus on the median rather than the average because the BF values within each prior specification tended to be positively skewed. This was expected, given that BFs can range from 0 to ∞. In this table, the prior specification of the mean intercept is presented in the rows, and the prior specification of the mean slope is presented in the columns. Table 3 shows that the BF value increased drastically as the degree of prior-data disagreement increased (i.e., as our eye moves towards the lower-right corner of the table, the values become larger).

Across sample sizes, we can see that, for $n = 50$, the BF appeared to increase if one or both priors were aligned and strongly informative. This pattern seems to mimic the patterns found for the DAC in that overly precise priors were penalized (to a certain extent). For $n = 100$ and 250, this effect became much less pronounced. For $n = 500$, the median BF value for prior specifications with aligned priors was often < 1, which indicates that the informative aligned priors appeared to result in a marginal likelihood that was more supportive of the data than the benchmark prior specification. For this sample size, the median BF value was also < 1 for several prior specifications that included diverging priors. Only once one or both priors were diverging and moderately or strongly informative did the median BF rapidly increase. With smaller sample sizes, all prior specifications that included at least one divergent prior quickly inflated the median BF to very large values. It should be noted that the variability in BF values was much larger for $n = 500$ compared to smaller sample sizes. For example, across diffuse and aligned priors, the average standard deviation of BF values was 5.14 for $n = 50$, 1.87 for $n = 100$, 0.69 for $n = 250$, but 29,664,081.00 for $n = 500$. This unexpected behavior will be further discussed in Section 4.3.4.4.

Table 5. Median BF value across simulation conditions.

Sample Size	Prior: Intercept	Prior: Slope						
		Diffuse	Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	1.00	1.19	1.60	2.74	5.39	67.03	899.77
	Aligned: Weak	1.17	1.41	1.91	3.28	6.71	85.50	1201.86
	Aligned: Moderate	1.56	1.92	2.71	4.68	9.68	136.09	2005.94
	Aligned: Strong	2.64	3.22	4.57	8.02	17.72	270.52	4019.66
	Diverged: Weak	5.58	6.78	9.89	17.58	38.23	659.19	1.07E+04
	Diverged: Moderate	86.85	106.23	160.37	316.81	796.09	18974.30	3.73E+05
	Diverged: Strong	1.58E+03	2.22E+03	3.49E+03	6.62E+03	1.91E+04	6.22E+05	1.12E+07
100	Diffuse	1.00	1.10	1.31	1.97	6.01	222.08	47268.44
	Aligned: Weak	1.08	1.19	1.46	2.19	6.58	246.71	54858.35
	Aligned: Moderate	1.30	1.46	1.79	2.75	8.41	315.20	79323.84
	Aligned: Strong	1.97	2.19	2.71	4.28	13.04	511.92	1.42E+05
	Diverged: Weak	6.28	7.00	8.79	13.88	42.45	1838.31	5.45E+05
	Diverged: Moderate	288.25	324.70	425.97	695.78	2368.55	1.61E+05	7.99E+07
	Diverged: Strong	1.05E+05	1.22E+05	1.73E+05	3.32E+05	1.21E+06	1.47E+08	9.62E+10
250	Diffuse	1.00	1.02	1.14	1.47	6.41	695.73	5.20E+07
	Aligned: Weak	1.01	1.07	1.20	1.57	6.65	695.67	5.69E+07
	Aligned: Moderate	1.14	1.20	1.33	1.73	7.36	815.52	6.76E+07
	Aligned: Strong	1.53	1.53	1.74	2.31	9.60	1067.62	9.75E+07
	Diverged: Weak	6.30	6.64	7.45	10.01	43.06	4845.42	4.23E+08
	Diverged: Moderate	742.66	778.50	868.82	1220.93	5586.46	8.58E+05	1.23E+11
	Diverged: Strong	8.78E+07	8.99E+07	1.03E+08	1.54E+08	7.85E+08	2.28E+11	1.32E+17
500	Diffuse	1.00	0.00	0.00	0.00	0.02	3.90	6.38E+06
	Aligned: Weak	0.01	0.00	0.00	0.00	0.00	0.02	82,451.96
	Aligned: Moderate	0.00	0.00	0.00	0.00	0.00	0.01	22,534.66
	Aligned: Strong	0.00	0.00	0.00	0.00	0.00	0.01	25,457.56
	Diverged: Weak	0.05	0.00	0.00	0.00	0.00	0.14	8.21E+05
	Diverged: Moderate	4.22	0.01	0.01	0.00	0.22	8.88	6.23E+07
	Diverged: Strong	1.89E+07	93,267.54	43,533.25	26,197.22	5.01E+05	7.32E+07	1.10E+15

Note. BFs > 100,000 are presented in scientific format.

4.3.4.2 Using a Cutoff Value for the BF

Based on the previous section, we can conclude that the BF tends to increase as the magnitude of prior-data disagreement increases. But, at what point is a BF high enough to represent convincing evidence of prior-data disagreement? Rules of thumb have been suggested in the past, such as BF > 3 or > 10 (Jeffreys, 1961; Kass & Raftery, 1995). However, research has shown that the BF depends on, for example, the sample size (Morey & Rouder, 2011). This effect was also visible in Table 5, where the BF increased much more drastically for $n = 250$ compared to $n = 100$ and $n = 50$. Thus, I used an analytical approach to finding an optimal cutoff value.

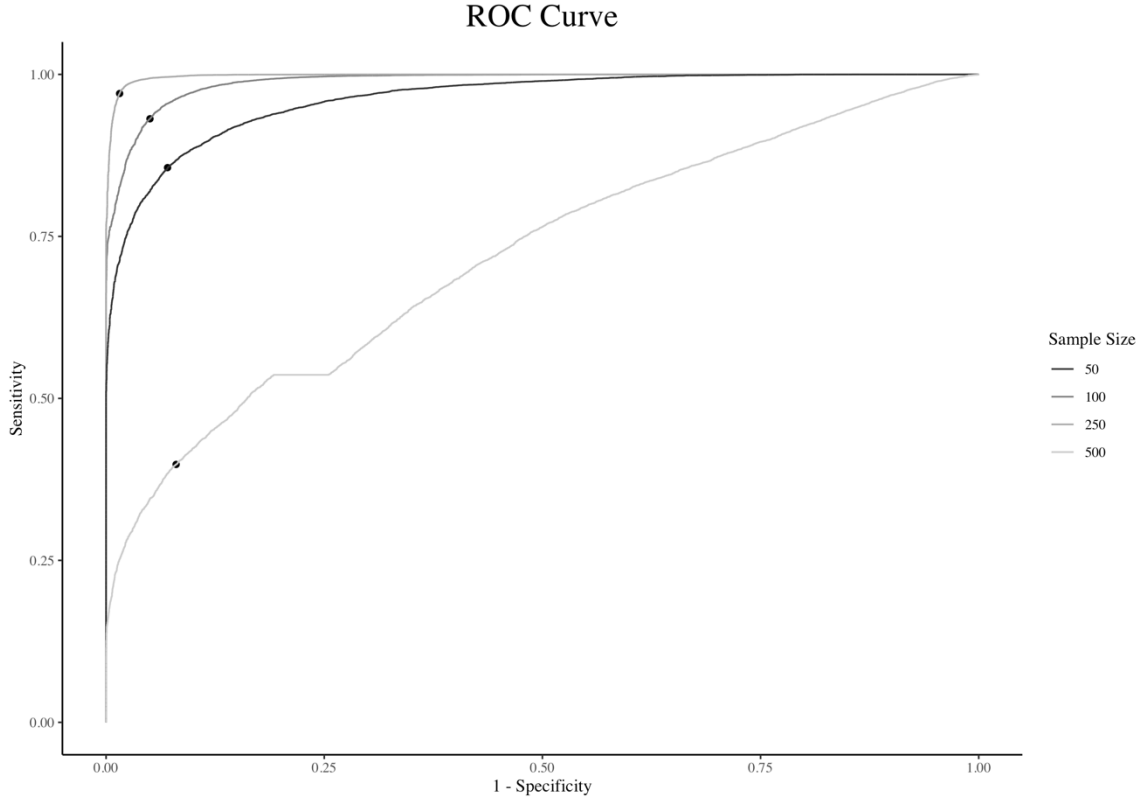


Figure 48. ROC curve of BF cutoff value located at optimal combination of sensitivity and specificity (the dots) for each sample size.

The results of this analysis are shown in Figure 48. The plot presents the receiver operating characteristic (ROC) curve for each sample size to illustrate the sensitivity, or true-positive rate (y-axis) and specificity, or false-positive rate (x-axis) of each cutoff value. This plot also illustrates that the selected cutoff value became more sensitive as the sample size increased from 50 to 250. The specific cutoff values were: 9.40 (90% interval = 9.02; 9.78, accuracy = 0.88) for $n = 50$, 5.67 (90% interval = 5.55; 5.81, accuracy = 0.94) for $n = 100$, and 4.20 (90% interval = 4.14; 4.28, accuracy = 0.98) for $n = 250$. Thus, the optimal cutoff value appeared to decrease as the sample size increased. In addition, decisions based on the cutoff value became more accurate (increasing from 0.88 to 0.98) as the sample size increased. However, for $n = 500$, a different picture emerged.

First, the selected cutoff value was much higher compared to the smaller sample size conditions: 346.30 (90% interval = 210.92; 760.86, accuracy = 0.57). Second, decisions based on that cutoff value became visibly less accurate. This inaccuracy is likely related to the increased variability in BF values across replications within prior specification conditions (see Section 4.3.4.1) and will be further discussed in Section 4.3.4.4.

Next, I applied these cutoff values to assess the proportion of times the BF indicated prior-data disagreement (Table 6). Ideally, these proportions should be 1 for all cells with at least one divergent moderately or strongly informative prior (the outer two columns and lower two rows within each sample size). In addition, the proportion should be zero for all cells with diffuse or aligned priors.

The table shows that, for n = 50 to 250, the BF cutoff values were likely to indicate prior-data disagreement when one or both priors were divergent and moderately or strongly informative. Furthermore, the performance of the cutoff values improved as the sample size increased. However, the table also shows that if one or both priors are aligned and strongly informative, the cutoff values became more likely to indicate prior-data disagreement. For example, if both priors were aligned and strongly informative, this proportion was .445 for n = 50, .341 for n = 100, and .095 for n = 250. This unexpected finding makes sense if one considers that, with a smaller sample size, one is more likely to encounter a strange sample, a sample that is not a good reflection of the underlying population. With informative priors that place the prior probability more heavily around the population values, a strange sample will stick out more and will likely result in prior-data disagreement.

For n = 500, the results shown in Table 6 further illustrate the lower sensitivity and specificity of the selected cutoff value (see Figure 48). In line with smaller sample sizes, divergent prior specifications were most likely to be flagged as prior-data disagreement, albeit at a lower rate. However, the BF cutoff value was also likely to indicate prior-data disagreement for prior specifications that included a combination of diffuse and aligned priors. Further, for n = 500, aligned weakly or moderately informative prior specifications sometimes resulted in BFs above the cutoff value. The performance of the BF cutoff value is markedly worse for n = 500 than for n = 250, for which fully aligned weakly or moderately informative prior specifications never resulted in BFs above the cutoff value.

Table 6. Proportion of times the BF indicated prior-data disagreement.

Sample Size	Prior: Intercept	Prior: Slope						
		Diffuse	Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	0.000	0.000	0.000	0.050	0.290	0.888	0.994
	Aligned: Weak	0.000	0.000	0.000	0.070	0.360	0.932	0.994
	Aligned: Moderate	0.002	0.010	0.034	0.160	0.498	0.960	0.998
	Aligned: Strong	0.050	0.064	0.164	0.412	0.752	0.986	1.000
	Diverged: Weak	0.258	0.338	0.522	0.804	0.934	1.000	1.000
	Diverged: Moderate	0.894	0.924	0.964	0.992	1.000	1.000	1.000
	Diverged: Strong	0.984	0.988	0.998	1.000	1.000	1.000	1.000
100	Diffuse	0.000	0.000	0.000	0.040	0.538	1.000	1.000
	Aligned: Weak	0.000	0.000	0.000	0.050	0.598	1.000	1.000
	Aligned: Moderate	0.002	0.002	0.006	0.084	0.738	1.000	1.000
	Aligned: Strong	0.056	0.066	0.102	0.282	0.936	1.000	1.000
	Diverged: Weak	0.548	0.606	0.730	0.932	0.996	1.000	1.000
	Diverged: Moderate	0.994	0.996	0.996	1.000	1.000	1.000	1.000
	Diverged: Strong	1.000	1.000	1.000	1.000	1.000	1.000	1.000
250	Diffuse	0.000	0.000	0.000	0.012	0.800	1.000	1.000
	Aligned: Weak	0.000	0.000	0.000	0.022	0.826	1.000	1.000
	Aligned: Moderate	0.000	0.000	0.000	0.038	0.892	1.000	1.000
	Aligned: Strong	0.020	0.024	0.036	0.110	0.966	1.000	1.000
	Diverged: Weak	0.832	0.862	0.906	0.956	1.000	1.000	1.000
	Diverged: Moderate	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Diverged: Strong	1.000	1.000	1.000	1.000	1.000	1.000	1.000
500	Diffuse	0.000	0.150	0.146	0.140	0.190	0.318	0.822
	Aligned: Weak	0.202	0.068	0.058	0.034	0.072	0.176	0.728
	Aligned: Moderate	0.158	0.068	0.044	0.042	0.086	0.150	0.678
	Aligned: Strong	0.144	0.054	0.032	0.022	0.070	0.154	0.676
	Diverged: Weak	0.192	0.084	0.082	0.080	0.122	0.246	0.760
	Diverged: Moderate	0.340	0.180	0.156	0.150	0.226	0.338	0.868
	Diverged: Strong	0.840	0.708	0.676	0.658	0.758	0.898	1.000

4.3.4.3 Optimal Prior as Identified by the BF

In addition to using cutoff values, it is also possible to compare across prior specifications to examine how often each prior specification results in the lowest BF. The prior specification with the smallest BF value reflects a prior most in agreement with the data. Table 7 presents the proportion of times each prior specification combination (across intercept and slope priors) resulted in the lowest BF value. Only specifications that were selected are included in the table. For $n = 50$, the smallest BF was most often associated with prior specifications that had diffuse priors on one or both parameters. As the sample size increased, this preference became less apparent. Instead, the lowest BF started to be associated with a prior specification that included aligned, weakly or moderately informative priors on one of both parameters. Models with divergent priors were never selected.

These findings indicate that, although the BF is unlikely to select a model with divergent priors among a set of prior specifications, it may not be the best prior-data disagreement index for selecting the *optimal* prior specification in terms of minimizing the relative bias and RMSE of the posterior estimates. As the choice of prior specification has a larger impact on the posterior parameter estimates with smaller samples, it is disappointing to observe that the BF is unlikely to prefer aligned informative priors over diffuse priors for the smallest sample size examined.

4.3.4.4 Issues with Computing the BF for Larger Sample Sizes

The previous subsections have illustrated that the BF appears to be unstable and unreliable for $n = 500$. Compared to smaller sample sizes, with $n = 500$, the BF was much more variable and less clearly related to the prior specification. One reason for this instability is that the marginal likelihood of this model (an LGM) is much more complex for $n = 500$ than for $n = 50$ or even 250. The core number of parameters estimated is the same across sample sizes and include five residual variances, the mean intercept and slope, and the covariance matrix of the intercept and slope. However, for each case in the sample, two additional parameters are estimated: the individual's intercept and slope. Thus, for $n = 50$, a total of 110 unique parameters are estimated. In contrast, for $n = 500$, a total of 1,010 parameters are estimated. The creators of the R package 'bridgesampling', used in the current study, suggest that for complex models, "testing requires about an order of magnitude more posterior samples than estimation" (Gronau et al., 2020 p. 12). For the current study, this would result in a total of 100,000 posterior samples for each of 49 prior specifications across 1,000 replications. Using the most powerful computer I have access to, estimating posteriors for 49 prior specifications for one replication would take about 50 hours to complete. This estimate assumes that the amount of memory available is sufficient to save all the posterior samples. Thus, even using all six simulation computers I have access to, it would take at least $50 \times 1000 / 24 / 6 \approx 348$ days, which is unfeasible. The implications of this finding will be further addressed in the Discussion section.

Table 7. Proportion of times a specific prior specification resulted in the lowest BF value.

Sample Size	Prior: Intercept	Prior: Slope						
		Diffuse	Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	**0.997**	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	**0.003**	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	Diffuse	**0.891**	**0.054**	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	**0.054**	**0.001**	0.000	0.000	0.000	0.000	0.000
	Aligned: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
250	Diffuse	**0.231**	**0.187**	**0.075**	**0.005**	0.000	0.000	0.000
	Aligned: Weak	**0.178**	**0.105**	**0.052**	**0.002**	0.000	0.000	0.000
	Aligned: Moderate	**0.080**	**0.047**	**0.025**	**0.001**	0.000	0.000	0.000
	Aligned: Strong	**0.004**	**0.008**	0.000	0.000	0.000	0.000	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	Diffuse	0.000	**0.001**	0.000	**0.002**	0.000	0.000	0.000
	Aligned: Weak	**0.001**	**0.039**	**0.078**	**0.080**	**0.013**	0.000	0.000
	Aligned: Moderate	0.000	**0.061**	**0.100**	**0.145**	**0.024**	0.000	0.000
	Aligned: Strong	**0.001**	**0.092**	**0.105**	**0.171**	**0.022**	0.000	0.000
	Diverged: Weak	**0.001**	**0.012**	**0.019**	**0.028**	**0.002**	0.000	0.000
	Diverged: Moderate	0.000	0.000	**0.002**	**0.001**	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. Proportions > 0 were bolded for emphasis.

4.3.5 The Prior-Predictive *p*-value

The motivation for estimating an LGM is that the latent growth parameters, such as the linear intercept and slope, provide a good representation of the underlying data at each time point. Thus, to examine whether the priors placed on the mean intercept and slope parameters are in line with the underlying observed data, the observed sample means at each time point were compared to the prior-predictive samples' prior means. A prior-predictive *p*-value < .05 or > .95 indicated that the prior specification resulted in prior-predictive samples that did not provide a good representation in terms of the average values at each time point. For the current population model, the first and last time point will likely be most affected by the prior specification of the intercept and slope mean, respectively. Thus, the results will focus on these two time points.

4.3.5.1 Prior-Predictive *p*-values that Indicate Poor Fit

The prior-predictive *p*-value for the mean of *y1* only exceeded the commonly used cutoff values of < .05 and >.95 for one prior specification with one sample size: a diffuse prior specified for the intercept mean and an aligned moderately informative prior for the slope mean with $n = 500$. Inspecting the prior-predictive samples for this specific prior specification revealed that the prior-predictive mean of *y1* was -30.94 ($SD = 18.70$). In contrast, the mean of *y1* across all 1,000 replications with $n = 500$ was 1.00 ($SD = 0.06$). As the aligned prior specified for the slope mean is not expected to affect the observations at the first time point, one may expect that the prior-predictive mean would be closer to zero, the center of the diffuse prior placed on the intercept. However, -30.94 is still within one standard deviation (i.e., 55) from the mean hyperparameter. A technical explanation of the potential issues of using prior-predictive samples with diffuse priors is provided in Appendix A.

Table 8 presents the proportion of times a prior-predictive *p*-value indicated poor fit based on the same cutoff values for *y5*. Only sample sizes for which at least one prior specification resulted in a proportion > 0 were included in the tables. Based on Table 9, the prior-predictive *p*-value exceeded the cutoff values in the two larger sample size conditions when looking at *y5.* For $n = 250$, the prior-predictive *p*-value indicated poor prior-data fit for some replications if the priors diverged for both parameters. For $n = 500$, the prior-predictive *p*-value became more likely to indicate poor prior-data fit when diverging priors were specified for both parameters, and to a lesser extent, when a diverging prior was specified for the slope mean parameter. If both priors were divergent and moderately or strongly informative, the prior-predictive *p*-value exceeded the cutoff values for all replications.

The difference in results when computing the prior-predictive *p*-value for the mean of *y1* or *y5* highlights the importance of assessing different aspects of the observed data through prior-predictive checks. It appears that the diverging priors examined in the current study did not result in overly unusual prior-predictive samples for the mean of *y1* (apart from the prior specification mentioned above). Yet, those same diverging priors did emerge as problematic for generating prior-predictive samples for the mean of *y*5. However, these problematic prior-predictive samples emerged only for the two larger

sample sizes examined, indicating that relying on the cutoff values is likely not helpful in detecting prior-data disagreement with smaller sample sizes.

Table 8. Proportion of prior-predictive *p*-values for the mean of *y5* that exceed cutoff values for prior-data disagreement for *n* = 250 and 500.

		Prior: Slope						
Sample Size	Prior: Intercept	Diffuse	Weak	Aligned Moderate	Strong	Weak	Divergent Moderate	Strong
250	Diffuse	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Divergent: Weak	0.000	0.000	0.000	0.000	**0.001**	**0.029**	**0.245**
	Divergent: Moderate	0.000	0.000	0.000	0.000	**0.023**	**0.098**	**0.353**
	Divergent: Strong	0.000	0.000	0.000	0.000	0.000	**0.121**	**0.361**
500	Diffuse	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	0.000	0.000	0.000	0.000	0.000	**0.004**	**0.005**
	Aligned: Moderate	0.000	0.000	0.000	0.000	0.000	**0.006**	**0.030**
	Aligned: Strong	0.000	0.000	0.000	0.000	0.000	**0.021**	**0.016**
	Divergent: Weak	0.000	0.000	0.000	0.000	**0.708**	**0.990**	**1.000**
	Divergent: Moderate	0.000	0.000	0.000	0.000	**0.740**	**1.000**	**1.000**
	Divergent: Strong	0.000	0.000	0.000	0.000	**0.487**	**1.000**	**1.000**

Note. Proportions > 0 were bolded for emphasis, proportions were equal to 0 for *n* = 50 and 100.

Table 9. Absolute difference between the ideal and observed prior-predictive *p*-value of *y1* across simulation conditions.

Sample Size	Prior: Intercept	Prior: Slope						
		Diffuse	Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	0.188	0.002	0.066	0.011	0.119	0.051	0.004
	Aligned: Weak	0.036	0.037	0.036	0.035	0.037	0.038	0.036
	Aligned: Moderate	0.039	0.038	0.040	0.040	0.040	0.043	0.038
	Aligned: Strong	0.038	0.040	0.041	0.041	0.040	0.040	0.041
	Diverged: Weak	0.161	0.162	0.158	0.157	0.155	0.152	0.158
	Diverged: Moderate	0.160	0.168	0.161	0.161	0.166	0.160	0.153
	Diverged: Strong	0.154	0.169	0.167	0.159	0.168	0.166	0.167
100	Diffuse	0.115	0.105	0.050	0.082	0.096	0.111	0.013
	Aligned: Weak	0.032	0.032	0.033	0.032	0.033	0.033	0.035
	Aligned: Moderate	0.030	0.036	0.037	0.037	0.037	0.037	0.037
	Aligned: Strong	0.040	0.038	0.039	0.037	0.039	0.038	0.038
	Diverged: Weak	0.196	0.171	0.202	0.196	0.200	0.200	0.200
	Diverged: Moderate	0.238	0.212	0.209	0.207	0.202	0.211	0.219
	Diverged: Strong	0.212	0.213	0.217	0.215	0.224	0.199	0.214
250	Diffuse	0.213	0.137	0.189	0.291	0.152	0.073	0.114
	Aligned: Weak	0.029	0.030	0.035	0.028	0.029	0.030	0.028
	Aligned: Moderate	0.032	0.033	0.034	0.033	0.034	0.034	0.035
	Aligned: Strong	0.035	0.037	0.037	0.037	0.038	0.038	0.037
	Diverged: Weak	0.269	0.270	0.259	0.267	0.268	0.270	0.263
	Diverged: Moderate	0.296	0.282	0.276	0.284	0.287	0.281	0.289
	Diverged: Strong	0.271	0.286	0.284	0.286	0.272	0.286	0.284
500	Diffuse	0.026	0.136	0.471	0.177	0.093	0.176	0.120
	Aligned: Weak	0.027	0.027	0.025	0.026	0.033	0.026	0.025
	Aligned: Moderate	0.032	0.033	0.031	0.031	0.036	0.033	0.032
	Aligned: Strong	0.037	0.035	0.036	0.039	0.034	0.035	0.034
	Diverged: Weak	0.345	0.319	0.305	0.328	0.312	0.318	0.309
	Diverged: Moderate	0.340	0.338	0.323	0.334	0.335	0.342	0.339
	Diverged: Strong	0.341	0.342	0.353	0.344	0.341	0.341	0.346

Table 10. Absolute difference between the ideal and observed prior-predictive *p*-value of *y5* across simulation conditions.

Sample Size	Prior: Intercept	Prior: Slope						
		Diffuse	Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	0.039	0.002	0.064	0.011	0.107	0.039	0.008
	Aligned: Weak	0.007	0.058	0.061	0.065	0.227	0.243	0.252
	Aligned: Moderate	0.001	0.059	0.063	0.069	0.236	0.251	0.257
	Aligned: Strong	0.049	0.059	0.065	0.069	0.234	0.250	0.244
	Diverged: Weak	0.026	0.131	0.140	0.148	0.309	0.318	0.320
	Diverged: Moderate	0.004	0.136	0.144	0.149	0.308	0.321	0.325
	Diverged: Strong	0.039	0.128	0.150	0.151	0.307	0.316	0.317
100	Diffuse	0.004	0.105	0.051	0.090	0.086	0.125	0.002
	Aligned: Weak	0.019	0.053	0.062	0.064	0.270	0.298	0.304
	Aligned: Moderate	0.206	0.056	0.065	0.070	0.286	0.300	0.311
	Aligned: Strong	0.156	0.056	0.061	0.070	0.291	0.311	0.312
	Diverged: Weak	0.107	0.120	0.167	0.174	0.361	0.370	0.380
	Diverged: Moderate	0.014	0.161	0.177	0.180	0.362	0.373	0.368
	Diverged: Strong	0.149	0.164	0.181	0.186	0.364	0.375	0.373
250	Diffuse	0.038	0.138	0.189	0.291	0.145	0.097	0.095
	Aligned: Weak	0.095	0.047	0.056	0.059	0.349	0.371	0.366
	Aligned: Moderate	0.183	0.046	0.058	0.064	0.346	0.375	0.380
	Aligned: Strong	0.072	0.047	0.062	0.065	0.352	0.382	0.391
	Diverged: Weak	0.072	0.199	0.226	0.241	0.417	0.432	0.443
	Diverged: Moderate	0.090	0.210	0.235	0.261	0.427	0.436	0.446
	Diverged: Strong	0.363	0.213	0.244	0.253	0.413	0.438	0.446
500	Diffuse	0.073	0.137	0.471	0.175	0.073	0.199	0.103
	Aligned: Weak	0.095	0.043	0.049	0.054	0.389	0.417	0.422
	Aligned: Moderate	0.002	0.039	0.054	0.062	0.388	0.423	0.431
	Aligned: Strong	0.015	0.045	0.057	0.063	0.394	0.427	0.427
	Diverged: Weak	0.154	0.232	0.285	0.312	0.454	0.466	0.469
	Diverged: Moderate	0.276	0.242	0.311	0.314	0.454	0.470	0.475
	Diverged: Strong	0.236	0.232	0.303	0.326	0.449	0.470	0.473

4.3.5.2 Distance from the Ideal Prior-Predictive *p*-value of 0.5

Table 9 and 10 above present the absolute average distance between the prior-predictive *p*-value and the ideal value of .5 (which reflects a scenario in which half the prior-predictive sample means lie above and below the observed sample mean) for *y1* and *y5* respectively. The tables show that the prior-predictive *p*-value generally moved away from the ideal as the prior specification became more divergent (i.e., within each sample size, as our eye moves towards the lower-right corner of the table, the values become larger). Furthermore, this pattern became more pronounced as the sample size increased. Note that the average distance from the ideal value was also quite large for some prior specifications that included one or two diffuse priors (see Appendix A).

Focusing on the prior-predictive *p*-value for the mean of *y1* (Table 10), it is not unexpected that a divergent slope prior did not appear to affect the distance from the ideal value (i.e., the value does not change meaningfully moving from left-to-right within a row of the table). It is not strange that this value was most affected by the mean intercept prior. In contrast, the prior-predictive *p*-value for the mean of *y5* (Table 11) was affected by both prior specifications and was visibly farther removed from the ideal value if diverging priors were specified for both parameters.

This finding reinforces the implication that it is important to look at the prior-predictive ability of a model in multiple ways (e.g., looking at multiple observed variables) to fully assess the impact of each prior.

4.3.5.3 Optimal Prior as Identified by Prior-Predictive *p*-value

Finally, we can look at how often the prior-predictive *p*-value is closest to the ideal 0.5 for each prior specification to examine which prior specification a researcher would likely select across all options (Table 11 and 12 for *y1* and *y5* respectively).

Focusing first on the results for *y1* (Table 11), the optimal prior based on the prior-predictive *p*-value depended on the sample size. For $n = 50$, the fully diffuse specification was most often picked. More importantly, specifications that included one or two divergent priors were hardly ever preferred. The prior specifications that included only diffuse and aligned priors were selected 0.89 of the time.

For $n = 100$ and 250, the specification with a diffuse intercept prior and an aligned, weakly informative slope prior was most often selected. Similar to the results for $n = 50$, prior specifications that included a divergent intercept were hardly ever preferred. However, specifications in which an aligned intercept prior was combined with a diverging slope prior became more likely to emerge as the optimal prior specification. Prior specifications that included only diffuse and aligned priors were selected 0.93 of the time for $n = 100$ and 0.88 of the time for $n = 250$.

For $n = 500$, the prior combinations most often selected shifted markedly to the specification with an aligned, weakly informative intercept prior and a divergent, moderately informative slope prior. Overall, the likelihood of selecting any prior specification was more evenly distributed over all prior specifications in which the intercept prior was diffuse or aligned. Prior specifications that included only diffuse and aligned priors were selected 0.31 of the time. Thus, as the sample size increased, prior specifications that included a diverging prior for the mean slope parameter were more

likely to emerge as the optimal prior specification when it came to generating prior-predictive samples that are in line with the mean of *y1*.

A different pattern emerged if the mean of *y*5 was the target of the prior-predictive *p*-value (Table 12). For $n = 50$ and 100, prior specifications with a diverging intercept or slope prior were associated with the optimal prior-predictive *p*-value for some replications. Most notably, for $n = 100$, the prior specification with a diffuse prior on the intercept mean and a diverging, strongly informative prior on the slope mean resulted in a prior-predictive *p*-value closest to 0.5 for a large majority of the replications. This unexpected finding is further discussed in Appendix A. However, for $n = 250$ and 500, the optimal prior-predictive *p*-value was associated only with prior specifications that included diffuse or aligned priors. This finding further demonstrates the importance of assessing the prior-predictive ability of a model for different aspects of the observed data. A researcher who only focused on *y1* might have selected a prior specification that included a diverging prior for the slope mean parameter, resulting in increased parameter estimate bias and RMSE. A researcher who looked at *y1* and *y5* would be unlikely to make the same decision. However, even if both prior-predictive *p*-values were examined, a researcher would still be unlikely to select highly informative priors, which may not be optimal for minimizing posterior parameter estimate bias and RMSE.

Table 11. Proportion each prior specification selected based on prior-predictive *p*-value for the mean of *y1* across sample sizes.

Sample Size	Prior: Intercept	Diffuse	Prior: Slope					
			Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	0.000	**0.610**	0.000	0.000	0.000	0.000	**0.210**
	Aligned: Weak	**0.029**	**0.001**	**0.006**	**0.008**	**0.001**	**0.021**	**0.004**
	Aligned: Moderate	**0.006**	**0.003**	**0.003**	**0.007**	0.000	**0.044**	**0.002**
	Aligned: Strong	**0.005**	**0.006**	0.000	**0.006**	**0.033**	**0.004**	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	Diffuse	0.000	0.000	0.000	0.000	0.000	0.000	**0.612**
	Aligned: Weak	**0.009**	**0.016**	**0.013**	**0.003**	**0.004**	**0.108**	**0.004**
	Aligned: Moderate	**0.151**	**0.003**	**0.006**	**0.003**	**0.007**	**0.014**	**0.002**
	Aligned: Strong	**0.009**	**0.011**	**0.016**	**0.001**	**0.006**	**0.006**	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
250	Diffuse	0.000	0.000	0.000	0.000	0.000	**0.020**	0.000
	Aligned: Weak	**0.130**	**0.396**	**0.321**	**0.032**	**0.003**	**0.004**	**0.010**
	Aligned: Moderate	**0.015**	0.000	**0.005**	**0.008**	**0.013**	**0.004**	**0.006**
	Aligned: Strong	**0.005**	**0.007**	**0.003**	**0.008**	**0.016**	**0.002**	**0.004**
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	Diffuse	**0.186**	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	**0.045**	**0.038**	**0.008**	**0.131**	**0.205**	**0.004**	**0.025**
	Aligned: Moderate	**0.014**	**0.055**	**0.021**	**0.030**	**0.133**	**0.018**	**0.008**
	Aligned: Strong	**0.012**	**0.013**	**0.006**	**0.014**	**0.018**	**0.002**	**0.015**
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. Proportions > 0 bolded for emphasis.

Table 12. Proportion each prior specification selected based on prior-predictive *p*-value for the mean of *y5* across sample sizes.

Sample Size	Prior: Intercept	Prior: Slope						
		Diffuse	Aligned			Diverged		
			Weak	Moderate	Strong	Weak	Moderate	Strong
50	Diffuse	0.000	**0.336**	0.000	**0.003**	0.000	0.000	**0.002**
	Aligned: Weak	0.000	**0.004**	**0.008**	**0.005**	0.000	0.000	0.000
	Aligned: Moderate	**0.635**	**0.002**	**0.006**	**0.002**	0.000	0.000	0.000
	Aligned: Strong	0.000	**0.006**	0.000	**0.003**	0.000	0.000	0.000
	Diverged: Weak	0.000	0.000	**0.001**	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	**0.001**	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
100	Diffuse	**0.087**	0.000	0.000	0.000	0.000	0.000	**0.899**
	Aligned: Weak	0.000	**0.002**	0.000	**0.001**	0.000	0.000	0.000
	Aligned: Moderate	0.000	**0.002**	**0.001**	**0.001**	0.000	0.000	0.000
	Aligned: Strong	0.000	**0.001**	0.000	0.000	0.000	0.000	0.000
	Diverged: Weak	0.000	**0.007**	0.000	**0.001**	0.000	0.000	0.000
	Diverged: Moderate	0.000	**0.001**	0.000	**0.001**	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	**0.001**	0.000	0.000	0.000	0.000
250	Diffuse	**0.430**	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	0.000	**0.255**	**0.193**	**0.032**	0.000	0.000	0.000
	Aligned: Moderate	0.000	**0.020**	**0.031**	**0.003**	0.000	0.000	0.000
	Aligned: Strong	0.000	**0.012**	**0.019**	**0.008**	0.000	0.000	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000
500	Diffuse	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Aligned: Weak	0.000	**0.029**	**0.004**	**0.019**	0.000	0.000	0.000
	Aligned: Moderate	**0.856**	**0.024**	**0.007**	**0.003**	0.000	0.000	0.000
	Aligned: Strong	0.000	**0.030**	**0.013**	**0.019**	0.000	0.000	0.000
	Diverged: Weak	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Moderate	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Diverged: Strong	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. Proportions > 0 bolded for emphasis.

4.4 Discussion

Study 2 examined the ability of the DAC, BF, and a prior-predictive checking procedure (i.e., prior-predictive *p*-value) to detect prior-data disagreement in Bayesian SEM. Gaining a better understanding of the extent to which those indices can be used for assessing the presence of prior-data disagreement will help applied researchers become aware of the potential impact their priors may have on inferences based on the posterior parameter estimates. In the Introduction, I linked each of the three prior-data disagreement indices to a specific question that they could answer about the potential disagreement between the prior and the data. I will now discuss what the results of this study can tell us about the extent to which each of the indices can be used to answer those questions in the context of SEM.

4.4.1 The DAC

The specific question addressed with the DAC is: *Is this prior a good representation of the information present in the data about parameter θ?* The results of this study show that the DAC can be used to answer this question quite well. Notably, highly informative divergent priors were consistently flagged as disagreeing with the data across sample sizes. Moderately informative divergent priors were also likely flagged as disagreeing with the data for $n = 50$, and consistently flagged as disagreeing with the data for $n = 100$ or larger. These results are encouraging, as those two diverging prior specifications resulted in the most severe posterior parameter estimate bias and RMSE across the included sample sizes. For $n = 50$, the DAC sometimes indicated prior-data disagreement for the aligned, strongly informative prior. This may reflect the DAC's ability to detect prior-data disagreement due to a conflict of information: The researcher-specified prior was more precise than the information from the data. However, it may also reflect the increased likelihood of drawing a sample that is not representative of the population if the sample size is small. Thus, with small sample sizes, the DAC may indicate prior-data disagreement that is due to a diverging prior or due to a diverging sample.

For that reason, it may be beneficial for researchers to compare multiple potential prior specifications to assess whether a specific prior may minimize the prior-data disagreement. The results of this study showed that if the DAC was compared across prior specifications, it was likely that one of the aligned prior specifications was selected as the optimal prior specification. The larger the sample size, the more likely it became that the strongly informative aligned prior distribution was selected as the optimal prior specification. The DAC's tendency to prefer informative, aligned priors may be beneficial for parameter estimation accuracy, as the results showed that informative, aligned prior specifications resulted in lower relative bias and RMSE across sample sizes.

Overall, the DAC was easily implemented to assess prior-data disagreement for SEM in this study. Since its computation only requires the Bayesian estimation of the benchmark prior model to form the reference posterior, $\pi^J(\theta|y)$, the computation time was extremely fast compared to the two other approaches examined in this study. The resulting DAC can be used to make a clear, binary decision regarding prior-data disagreement based on a meaningful, prespecified cutoff. Furthermore, the presence of

prior-data disagreement can be examined for each parameter in isolation, making it easier for the researcher to identify priors for specific parameters that disagree with the data. However, this may also pose a potential drawback for the use of the DAC, as priors placed on different parameters may interact in unexpected ways (Depaoli et al., 2020). This interaction among priors cannot be captured by the DAC.

Furthermore, the value of the DAC depends on the proper specification of the benchmark priors and reference posterior. The large number of parameters often involved in SEM further complicates this process, as diffuse priors placed on some parameters may act in an informative manner when combined with diffuse priors placed on other parameters (Depaoli, 2013; Depaoli & Clifton, 2015; Smid & Winter, 2020; van Erp et al., 2018). For the current study, I examined several benchmark prior specifications (see Footnote 15) to find a specification that minimally impacted the posterior distributions of the mean intercept and slope parameters. Default priors that are specified in Bayesian estimation packages such as 'blavaan' (Merkle & Rosseel, 2018) may not be appropriate for estimating the reference posterior used to compute the DAC. Benchmark priors that are used for one study or model may not be appropriate for use with a different sample or model. Thus, using the DAC necessitates the thoughtful specification of a set of priors that have a minimal impact on the reference posteriors. Researchers may also examine the impact of their chosen benchmark priors through a sensitivity analysis of several alternative benchmark priors.

4.4.2 The BF

The question addressed with the BF is: *Does this prior result in a marginal likelihood that is more supportive of the data than a benchmark prior?* The results of the current study appear to indicate that the BF may be able to answer this question under certain circumstances. The BF was likely to detect diverging, strongly informative priors, even for the smallest sample size included in the current study. If both priors were diverging, evidence for prior-data disagreement was consistently detected, even if the priors were only weakly informative (for n = 50 to 250). This highlights an advantage of the BF over the DAC in that the BF considers all specified priors simultaneously. Whereas weakly informative, divergent priors were unlikely to be detected through the DAC, the BF values of models with two diverging, weakly informative priors tended to be visibly greater in magnitude compared to specifications with just one diverging weakly informative prior.

However, the BF tended to indicate prior-data disagreement if aligned, strongly informative prior specifications were used with smaller sample sizes (i.e., n = 50 or 100). This finding was unexpected, as previous research showed that only the DAC tended to penalize overly informative priors (Veen et al., 2018). The current study differs from previous research in that priors for two parameters were varied simultaneously. If only one of the two priors was aligned and strongly informative, the BF was generally unlikely to indicate prior-data disagreement. Only when both priors were aligned and strongly informative did the BF increase visibly. Thus, it may be that the threshold at which the BF starts penalizing a conflict of information is higher (i.e., multiple priors need to be overly informative), or that this tendency to penalize overly informative priors only emerges in more complex models (i.e., SEMs).

Moreover, while the DAC has a meaningful cutoff value that differentiates between prior-data agreement and disagreement (DAC > 1), the BF does not have a uniformly applicable cutoff value that can easily be applied. The results reported in the current study regarding the optimal cutoff value show that this value may decline as the sample size increases (at least up to $n = 250$). This finding is in line with previous research showing that a meaningful BF cutoff value for deciding between two models depends on the sample size (Morey & Rouder, 2011). While it was possible to use an analytical approach to find the optimal cutoff value in the current study, these cutoff values cannot be generalized beyond the population model and sample sizes examined. Although the results presented here showed that the BF rapidly increased in the presence of severe prior-data disagreement, it was less straightforward to differentiate between for example diverging weakly informative priors and aligned strongly informative priors. Thus, it may be difficult for applied researchers to discern if their prior specification reflects prior-data disagreement for their particular sample and model specification. For that reason, it may be more meaningful to compare BFs across multiple prior specifications and find the prior specification with the lowest BF. However, with smaller sample sizes, this approach is likely to prefer prior specifications that are less informative. As posterior estimates were less biased with aligned informative priors, the BFs tendency to prefer diffuse or aligned weakly informative priors is not ideal.

Furthermore, given the problems that arose with the largest sample size included in the study ($n = 500$), using the BF for assessing prior-data disagreement may be problematic for applied researchers using SEMs or other complex models with larger samples. A researcher who collects longitudinal data from 500 participants may think that the prior specification no longer affects the posterior estimates. However, the results of the current study indicated that with $n = 500$, the posterior estimates of the mean intercept and slope parameters were still meaningfully biased for some diverging prior specifications. Thus, assessing prior-data disagreement remains relevant even with these larger sample sizes. Although it is unlikely that an applied researcher would compare 49 prior specifications as was done in the current study, the time commitment becomes impractical even with fewer options. For example, if a researcher would like to compare six prior specifications for a sample of $n = 500$, it might take at least seven hours even if they have access to state-of-the-art simulation computers. Future research could explore techniques for optimizing the efficiency of posterior sampling of complex models such as SEMs so that comparing the marginal likelihood across several models or prior specifications becomes more feasible with larger sample sizes.

Finally, if the BF is used to compare a set of informative prior specifications to a benchmark (diffuse) prior specification, researchers are required to find a set of priors that minimally affect marginal likelihood. However, with the BF, researchers can also use an alternative approach, not examined in the current study. Specifically, researchers can directly compare the marginal likelihood generated through two different informative prior specifications. For an example of this process, I refer to Veen and colleagues (2018).

4.4.3 The prior-predictive *p*-value

The specific question that prior-predictive checks address is: *Is this prior able to predict my observed data well?* Or: *Are my observed data unexpected under this prior specification?* The results of the current study indicate that whether the prior-predictive *p*-value can answer these questions depends on several factors.

First, the ability to detect prior-data disagreement with the prior-predictive *p*-value depended to a large extent on the observed data that were at the center of the calculating of this *p*-value. If the observed mean of *y1* was used as the basis of the discrepancy function, prior-data disagreement was never detected using commonly used cutoff values. Moreover, although the prior-predictive *p*-value was sensitive to some extent to a diverging intercept mean prior, its value was barely affected by a diverging slope mean prior. This lack of impact reflects the way the LGM was specified, as the path between the slope mean parameter and the observed variable *y1* was fixed to 0 (see Figure 41). This model specification meant that prior specifications that included a diverging slope prior were sometimes selected as the optimal prior specification. In contrast, if the observed mean of *y5* was used as the basis of the discrepancy function, prior-data disagreement was more likely to be detected with larger samples (n = 250 and 500), particularly if both priors were diverging. Across sample sizes, the prior-predictive *p*-value was affected by a diverging prior specified for either parameter. Furthermore, the prior-predictive *p*-value was more strongly affected if both priors were diverging. That meant that prior specifications with diverging slope priors were never selected as the optimal prior specification. The ability to consider priors specified for all parameters simultaneously (as with the BF) is a clear advantage of the prior-predictive *p*-value over the DAC. Based on these findings, applied researchers are urged to examine multiple aspects of their observed data to thoroughly investigate to what extent their observed data are unexpected under their prior specification.

The above findings highlight the utility of the prior-predictive *p*-value based on prior-predictive samples generated with the R package ‘rstan’ (Stan Development Team, 2020). However, researchers should also be cautioned about using the prior-predictive *p*-value if they assess prior specifications for SEMs that are entirely or partially diffuse. As pointed out in the Results section and further discussed in Appendix A, factors unrelated to the data, model, or prior specification may affect the prior-predictive *p*-value if diffuse priors are specified. These factors include aspects of the computer used to estimate the prior-predictive samples (e.g., operating system, hardware, underlying libraries, and specific C++ compiler) and the random seed specified in the sampling() function of the ‘rstan’ package. Although diffuse priors may be unlikely to reflect prior-data disagreement, they may still affect the posterior samples in unexpected ways (Depaoli, 2013; Depaoli & Clifton, 2015; Smid & Winter, 2020; van Erp et al., 2018). For that reason, researchers may still want to examine whether their diffuse prior specification can predict their observed data well. I strongly urge these researchers to interpret the prior-predictive *p*-value with caution. Where possible, alternative random seeds (and when possible, computer setups) should be examined to assess the stability of the prior-predictive samples.

Finally, it should be noted that, compared to the other two indices examined in the current study, the findings regarding the prior-predictive *p*-value are the least generalizable, as they may depend on my choice of the minimal sufficient statistic (the observed variables' sample means). This implementation of the prior-predictive *p*-value is not invariant to the choice of minimal sufficient statistic (Jang, 2010). A potential solution that may be investigated in future research was proposed by Nott and colleagues (Nott et al., 2016). In their approach, the discrepancy function is defined by using the distance between the researcher specified prior, $\pi(\theta)$, and resulting posterior $\pi(\theta|y)$. This distance could be quantified through a distance measure such as the KL divergence and used to define the prior-predictive *p*-value with the following discrepancy function:

$$P_T = P(KL[(\theta|y)||\pi(\theta)] \leq KL[\pi(\theta|y^{sim})||\pi(\theta)]). \quad (39)$$

The advantage of this implementation of the prior-predictive *p*-value is that it depends on the observed data only through the posterior distribution. This means that the discrepancy function is a function of any sufficient statistic and thus invariant (Lek & van de Schoot, 2019; Nott et al., 2016). This adjustment would result in a prior-predictive *p*-value that is more similar to the DAC, but that does not rely on the specification of a benchmark prior or reference posterior. However, this implementation of the prior-predictive *p*-value addresses a different question than the prior-predictive *p*-value examined in the current study: *Is this prior able to predict my posterior well?* Or: *Is my posterior unexpected under this prior specification?* Future research may compare this approach to implementing the prior-predictive *p*-value that more directly reflects the level of agreement between the prior and some aspect of the observed data.

4.4.4 How Prior-Data Disagreement Relates to Prior Sensitivity Analysis

The assessment of prior-data disagreement may be considered complementary to a prior sensitivity analysis. A prior sensitivity analysis is done after a researcher estimates a model based on their prior specification. It allows the researcher to assess the impact of their prior specification on the posterior estimates as compared to those obtained using different priors (Depaoli et al., 2020; B. O. Muthén & Asparouhov, 2012). The results of a prior sensitivity analysis illustrate how robust the final model results (based on the priors that the researcher originally specified) are to different prior specifications. The three indices examined in the current study do not directly assess the impact of the prior specification on the posterior estimates but instead, assess the extent to which the priors align with the information provided through the observed data. Thus, they could be used as a diagnostic tool to evaluate why the final model results are or are not robust to the chosen prior specification. Recall that the presence of prior-data disagreement does not automatically imply that the prior was wrong. It may be just as likely that the observed data is unusual.

Furthermore, although not the focus of the current study, it should be noted that the prior-predictive samples that are generated to compute the prior-predictive *p*-value do not rely on the existence of any observed data. For that reason, these prior-predictive samples

form an important tool for assessing the appropriateness of the intended prior specification before any data are observed. A clear indication of future prior-data disagreement arises when the prior predictive samples based on a certain prior specification cover a range of values that are not expected to be observed under the data generating process. If this happens, a researcher will have to consider whether this discrepancy is due to an inappropriate prior or an inaccurate model (Evans & Moshonov, 2006b).

A concern that arises with a prior sensitivity analysis is the temptation to engage in questionable research practices to find the "best" results, also known as Bayesian HARKing (hypothesizing after results are known; Kerr, 1998). After comparing different priors, a researcher may discover that some alternative prior specification leads to model results that are more in line with their hypotheses. Once a researcher has observed this more favorable result, they may be tempted to switch to that alternative prior specification and pretend that it was in line with their prior belief all along. Similar to conducting a prior sensitivity analysis, checking for prior-data disagreement opens the door to HARKing. To minimize the temptation of fishing for the "best" prior specification, the scientific community should continue to push for transparent and reproducible reporting of research designs and analyses (Rouder et al., 2019; Shrout & Rodgers, 2018). I also want to urge researchers to remember that their chosen prior is just one factor that could lead to prior-data disagreement. It is my hope that researchers will take this new source of information and use it to make more informed decisions about all components involved in the analysis: the prior, the data, and the model.

4.4.5 Conclusions

Previous research has demonstrated that the prior specification can meaningfully affect inferences drawn based on SEMs (e.g., Depaoli, 2014; Depaoli et al., 2017; Finch & Miller, 2019; Holtmann et al., 2016; Marcoulides, 2018; Miočević et al., 2020; Dingjing Shi & Tong, 2017; Smid, Depaoli, et al., 2019). The assessment of prior-data disagreement helps researchers become aware of the potential impact of their prior specification. Alternatively, the assessment of prior-data disagreement may point out something unusual about the collected sample data. The current study illustrated to what extent three indices for detecting prior-data disagreement can be used to assess different aspects of this potential disagreement in SEMs.

Based on the results of Study 2, I urge researchers to compare multiple prior specifications, as this will shed more light on the presence of prior-data disagreement than the assessment of a single prior specification. While the DAC has advantages such as its ease of implementation, researchers should ensure that the benchmark prior minimally affects the reference posterior. Furthermore, if priors for multiple parameters are altered across prior specifications, the DAC cannot be used to assess interactions between priors within a specification. For that reason, the BF and prior-predictive *p*-value may be preferred. However, these indices come with their own drawbacks. Assessing prior-data disagreement through the BF may become unfeasible as model complexity, sample size, or the number of prior specifications increase. Use of the prior-predictive *p*-value may not be appropriate for prior specifications that are partially or fully diffuse. Moreover, the prior-predictive ability of each prior specification should be assessed for

different aspects of the observed data to understand the full impact of the prior specification. A limitation of all three indices is that they were unlikely to point out the aligned, strongly informative prior specification as the optimal choice, particularly with smaller samples, even though this prior specification resulted in the lowest bias and RMSE of the posterior estimates.

Chapter 5

General Discussion

In this section, I discuss the main implications and contributions of the two studies in this dissertation. I also discuss relevant limitations to the research presented in this dissertation and provide suggestions for future directions for methodological research.

5.1.1 Contributions

The primary purpose of this dissertation was to improve the applicability of Bayesian estimation of SEM for applied researchers. Through two simulation studies, I examined two components that are central to Bayesian estimation in general and Bayesian SEM in particular: model and prior specification. The findings of these two studies contribute to existing knowledge on these two topics by highlighting the strengths and drawbacks of methods available to researchers across a variety of population models and conditions.

An important finding of this dissertation was that comparing across multiple specifications often resulted in more accurate decisions, whether the focus was on multiple models (Study 1) or multiple priors (Study 2). For instance, while the Bayesian approximate fit indices often indicated good model fit for misspecified models, their value did decrease when a correctly specified model was compared to a misspecified model. Similarly, whereas diverging prior specifications did not necessarily result in a DAC or a prior-predictive *p*-value that indicated prior-data disagreement, these diverging priors were unlikely to be selected as the optimal prior specification once they were compared to aligned or diffuse priors. It is encouraging to find this pattern across both studies. Regarding statistical model specification, it is generally agreed that "[a]ll models are wrong but some are useful" (Box, 1979, p. 202) or "[e]very scientist in the back of his mind takes it for granted that even the best theory is likely to be an approximation to the true state of affairs" (Meehl, 2009, p. 113). The practice of model comparison allows researchers to identify a model that is least wrong, or most useful. We could similarly assume that all priors are wrong, but some are useful. Even with extensive prior knowledge, researchers are unlikely to specify a prior that is exactly in line with the underlying population parameter distribution. Here again, comparing different priors in terms of prior-data disagreement may provide information to the researcher about which prior is least wrong, at least compared to the observed data. As stated in Chapter 4, this kind of information may serve as a useful diagnostic tool to understand the results of a prior sensitivity analysis.

The assessment of prior-data disagreement may serve an additional purpose in Bayesian SEM. Study 1 revealed that diverging priors confounded the association between the model specification and decisions regarding model fit or selection. Correctly specified models were more likely to result in model fit index values indicative of poor fit. Conversely, aligned priors did not affect this association between model specification and model fit or selection. Researchers who include informative priors in their analysis

gain a deeper understanding of what drives their model fit and selection decisions by using the prior-data disagreement indices examined in Study 2. Following the Bayesian perspective, finding disagreement between the priors and the data does not imply that the priors are wrong or need to be altered (Spiegelhalter et al., 2014; Veen et al., 2018; Young & Pettit, 1996). However, being aware of a disagreement provides a context around a study's findings regarding the reported model fit and posterior estimates. If the priors align with the data, a researcher can be more confident that their model fit and selection assessment reflects the appropriateness of their model specification and not the disagreement between the priors and the data.

5.1.2 Limitations and Future Directions

Several limitations emerged in the process of completing this dissertation that all provide avenues for future research to explore. For Study 1, I relied on commonly used cutoff values, such as BCFI > .95 or PPP < .05. While this design choice was made to be in line with the general use of these indices by applied researchers, numerous studies, most within the frequentist framework, have shown that cutoff values do not necessarily generalize across models, sample sizes, and other aspects of an analysis (e.g., Leite & Stapleton, 2006; Mcneish & Hancock, 2018; Niemand & Mai, 2018; Xia & Yang, 2018). One important future direction is to provide further insight into the sensitivity of the Bayesian approximate fit indices to factors other than model specification and provide guidance as to when cutoff values may be appropriate.

Another opportunity for future research is to compare the impact of different missing data mechanisms on the ability of the Bayesian model fit indices to detect misspecification. In Study 1, I focused on data that were MAR, or ignorable. Similarly, Asparouhov and Muthén (Asparouhov & Muthén, 2020) examined data that were MCAR and MAR, both considered ignorable. However, data collected for social science research may be more likely to include data that are MNAR, a nonignorable type of missing data. For example, in an educational research testing scenario with multiple items on an assessment, values may be missing because participants did not reach one or more items (Rose et al., 2017). Similarly, in survey research in which not all questions are mandatory, participants may choose not to answer certain items, based on item content or some other factor (C. W. Liu & Wang, 2017). Dropout from a longitudinal study can also be caused by a MNAR mechanism (Cuer et al., 2020). As these mechanisms may not always be apparent, it is essential to examine the consequences of assuming that data are MAR when they are in fact MNAR. In addition, it may also be interesting to examine if explicitly modeling the MNAR mechanism improves model fit assessment.

As stated in Chapter 4 and Appendix A, I discovered two major limitations of the prior-data disagreement indices investigated in Study 2. First, the computation of the BF became untenable as model complexity and sample size increased. Second, the generation of the prior-predictive samples was unstable for prior specifications that were entirely or partially diffuse. Whereas the first issue may be addressed through the development of more efficient estimation methods for the marginal likelihood or the introduction of ever more powerful computers, the second issue is more challenging to address. Researchers interested in examining prior-predictive samples generated from primarily diffuse prior specifications will need to interpret their findings with caution.

Finally, the conclusions drawn from the results of the studies in this dissertation are limited to the conditions included in the simulation designs. While the simulations were designed to represent realistic research scenarios, it is impossible to incorporate all the nuances that a researcher may encounter. Future research may identify other factors that affect model fit and selection assessment, such as measurement quality (Beauducel & Wittmann, 2005; Heene et al., 2011; Mcneish et al., 2018), skewness or kurtosis in the observed variables (Maydeu-Olivares et al., 2017), or categorical variables (Garrido et al., 2016). In addition, researchers might focus on other SEMs that benefit from Bayesian estimation, such as growth mixture models (Depaoli et al., 2019).

5.1.2.1 Recommendations for Applied Researchers

In this section, I lay out practical recommendations for applied researchers based on the findings of Study 1 and Study 2. For researchers interested in assessing model fit for a single model, I recommend focusing on the PPP-value as opposed to the approximate fit indices. With small sample sizes (and complete data), the approximate model fit indices may falsely indicate that a correctly specified model is misspecified, whereas the PPP-value is unlikely to falsely reject a correctly specified model, even with small sample sizes and missing values. However, if a researcher only has access to a relatively small sample (i.e., $n \leq 100$), a PPP-value > .05 does not necessarily indicate that a model is correctly specified, particularly if the data contain many missing values. Under these conditions, even a meaningful misspecification such as the omission of a quadratic slope is unlikely to cross any of the model fit indices' cutoff values (including the approximate fit indices). A researcher who has access to a larger sample (e.g., $n = 500$) without missing values, may compare their conclusion based on the PPP-value to their conclusion based on the approximate fit indices. Here, the BRMSEA appears more sensitive to misspecification than the BCFI and BTLI. If both indicate that a model is misspecified, the magnitude of the misspecification is likely substantial. In contrast, if only the PPP-value indicates that the model is misspecified, the misspecification may be more subtle: the model fits approximately, but not absolutely, well. Finally, researchers whose data include a large number of missing values across the majority of the observed values may not be able to interpret the approximate fit indices at all, as they cannot be computed. Under these circumstances, researchers will have to rely on the PPP-value, which will be somewhat less likely to detect misspecification (although the overall sample size is more important to this ability than the number of missing values).

For researchers interested in comparing multiple models and selecting the best specification, I recommend focusing on the approximate model fit indices over the PPP-value, BIC, and DIC. Even with small sample sizes, model selection based on the approximate model fit indices is more likely to result in selecting the correctly specified model. However, the approximate fit indices may be equivalent across model specifications if the sample size is small (i.e., $n \leq 100$) or if missing values are prevalent. Under those circumstances, the PPP-value can serve as an alternative. Compared to the BIC and DIC, the PPP-value is more likely to select the correctly specified model. The BIC and, to a lesser extent, the DIC are more likely to prefer a parsimonious, slightly misspecified model over the correctly specified model.

Researchers should be careful to interpret model fit indices if they have specified informative priors on (some of) the parameters. If the prior specification disagrees with the information provided by the data, model fit indices may falsely reject the correctly specified model. This is particularly true for the BRMSEA, which still appears to be affected by diverging priors when $n = 250$. To gain a better understanding of how the informative priors may have affected the model fit indices, researchers should assess the potential presence of prior-data disagreement.

Researchers can use multiple approaches to assess prior-data disagreement. Independent of the approach, I recommend that researchers examine multiple prior specifications, in line with a prior sensitivity analysis. Alternative prior specifications may include priors that are centered on the same value as the original prior specification, but that convey less precise prior knowledge (e.g., by increasing the variance hyperparameter). In addition, if researchers want to use the DAC or BF approach to assessing prior-data disagreement, they will need to use an appropriate diffuse prior specification. If researchers are unable to specify a diffuse prior specification (e.g., because diffuse priors cause convergence issues), they can examine to what extent their observed data are unexpected under their prior specification with the prior-predictive *p*-value. I do not recommend that use of the prior predictive *p*-value for diffuse prior specifications, as the results may depend on the specific computer setup used to generate the prior-predictive samples.

For researchers who can use an appropriate diffuse prior specification, the DAC is likely to flag meaningful prior-data disagreement across sample sizes. However, the DAC is computed for each prior-parameter pair sequentially, which may become impractical when models include many parameters. In that case, the BF may be used as an alternative, as long as the sample size is not too large, and the model is relatively simple (in terms of number of parameters). An advantage of the BF over the DAC is that entire prior specifications can be compared, which may reveal that some priors interact to result in prior-data disagreement. Comparing multiple prior specifications is essential here, as the BF does not have an intrinsic cutoff value that generalizes across sample sizes (or model complexities).

Overall, it is my hope that the results of these two studies can serve as a starting point for researchers as they embark on their next Bayesian SEM adventure.

References

Ainur, A. K., Sayang, M. D., Jannoo, Z., & Yap, B. W. (2017). Sample size and non-normality effects on goodness of fit measures in structural equation models. *Pertanika Journal of Science and Technology*, *25*(2), 575–586.

Asparouhov, T., & Muthén, B. (2010a). *Bayesian Analysis of Latent Variable Models using Mplus*. https://www.statmodel.com/download/BayesAdvantages18.pdf

Asparouhov, T., & Muthén, B. (2010b). *Bayesian Analysis Using Mplus: Technical Implementation*. https://www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., & Muthén, B. (2019). *Advances in Bayesian Model Fit Evaluation for Structural Equation Models*.

Asparouhov, T., & Muthén, B. (2020). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2020.1764360

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*(6), 1561–1577. https://doi.org/10.1177/0149206315591075

Baldwin, S. A., & Fellingham, G. W. (2012). Bayesian Methods for the Analysis of Small Sample Multilevel Data With a Complex Variance Structure. *Psychological Methods*, *18*(2), 151–164. https://doi.org/10.1037/a0030642

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338–363. https://doi.org/10.1037/1082-989X.8.3.338

Bayes, T. (1764). An Essay toward solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Beauducel, A., & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure. *Structural Equation Modeling*, *12*(1), 41–75. https://doi.org/10.1207/s15328007sem1201_3

Bell, M. L., Fiero, M., Horton, N. J., & Hsu, C. H. (2014). Handling missing data in RCTs; A review of the top medical journals. *BMC Medical Research Methodology*, *14*(1), 1–8. https://doi.org/10.1186/1471-2288-14-118

Bentler, P. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, *107*(2), 238–246.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. https://doi.org/10.1214/06-BA115

Betancourt, M. (2018). *A Conceptual Introduction to Hamiltonian Monte Carlo*.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.), *Current Trends in Bayesian Methodology with Applications* (pp. 79–101). CRC Press. https://doi.org/10.1201/b18502-5

Betancourt, M., & Stein, L. C. (2011). *The Geometry of Hamiltonian Monte Carlo*.

Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, *35*(9), 1011–1029. https://doi.org/10.1080/02664760802192981

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In R. Launer & G. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). Academic Press. https://doi.org/10.1016/b978-0-12-438150-6.50018-2

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)*, *143*(4), 383. https://doi.org/10.2307/2982063

Box, G. E. P., & Tiao, G. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley.

Brooks, S P, & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J Comp Graph Stat*, *7*(4), 434–455.

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An Evaluation of PPP and DIC for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 39–50. https://doi.org/10.1080/10705511.2018.1490648

Celeux, G., Forbesy, F., Robertz, C. P., & Titteringtonx, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(4), 651–674. https://doi.org/10.1214/06-BA122

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cuer, B., Mollevi, C., Anota, A., Charton, E., Juzyna, B., Conroy, T., & Touraine, C. (2020). Handling informative dropout in longitudinal analysis of health-related quality of life: Application of three approaches to data from the esophageal cancer clinical trial PRODIGE 5/ACCORD 17. *BMC Medical Research Methodology*, *20*(1), 1–13. https://doi.org/10.1186/s12874-020-01104-w

de la Horra, J., & Rodriguez-Bernal, M. T. (2003). Bayesian Robustness of the Posterior Predictive p-Value. *Communications in Statistics - Theory and Methods*, *32*(8), 1493–1503. https://doi.org/10.1081/STA-120022241

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . In *Journal of the Royal Statistical Society: Series B (Methodological)* (Vol. 39, Issue 1). https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186–219. https://doi.org/10.1037/a0031609

Depaoli, S. (2014). The Impact of Inaccurate "Informative" Priors for Growth Parameters in Bayesian Growth Mixture Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 239–252. https://doi.org/10.1080/10705511.2014.882686

Depaoli, S., & Clifton, J. P. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes,. *Structural Equation Modeling: A Multidisciplinary Journal* , *22*(3), 327–351.

https://doi.org/10.1080/10705511.2014.937849
Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065
Depaoli, S., Winter, S. D., Lai, K., & Guerra-Peña, K. (2019). Implementing continuous non-normal skewed distributions in latent growth mixture modeling: An assessment of specification errors and class enumeration. *Multivariate Behavioral Research*, *54*(6), 795–821. https://doi.org/10.1080/00273171.2019.1593813
Depaoli, S., Winter, S. D., & Visser, M. (2020). The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. *Frontiers in Psychology*, *11*(November), 1–18. https://doi.org/10.3389/fpsyg.2020.608045
Depaoli, S., Yang, Y., & Felt, J. (2017). Using Bayesian Statistics to Model Uncertainty in Mixture Models: A Sensitivity Analysis of Priors. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 198–215. https://doi.org/10.1080/10705511.2016.1250640
Dwirifqi Kharisma Putra, M., Rahayu, W., & Umar, J. (2019). Indonesian-language version of general self-efficacy scale-12 using Bayesian confirmatory factor analysis: a construct validity testing. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *23*(1), 12–25. https://doi.org/10.21831/pep.v23i1.20008
Dyk, D. A. V., & Meng, X. L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1–50. https://doi.org/10.1198/10618600152418584
Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.
Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*(1), 76–93. https://doi.org/10.1037/met0000102
Evans, M., & Jang, G. H. (2010). *A Limit Result for the Prior Predictive. Technical Report No. 1004*.
Evans, M., & Moshonov, H. (2006a). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4), 893–914. https://doi.org/10.1214/06-BA129
Evans, M., & Moshonov, H. (2006b). Checking for Prior-Data Conflict. *Bayesian Analysis*, *1*(4), 893–914. http://genealogy.math.ndsu.nodak.edu/html/id.phtml?id=15995
Fan, X., & Sivo, S. A. (2005). Sensitivity of Fit Indexes to Misspecified Structural or Measurement Model Components: Rationale of Two-Index Strategy Revisited. *Structural Equation Modeling*, *12*(3), 343–367. https://doi.org/10.1207/s15328007sem1203_1
Fan, X., & Sivo, S. A. (2007). Sensitivity of Fit Indices to Model Misspecification and Model Types. *Multivariate Behavioral Research*, *42*(3), 509–529. https://doi.org/10.1080/00273170701382864
Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*(1), 56–83. https://doi.org/10.1080/10705519909540119

Fay, D. M., Levy, R., & Schulte, A. C. (2020). Model Criticism of Growth Curve Models via Posterior Predictive Model Checking Model Criticism of Growth Curve Models via Posterior Predictive Model Checking. *The Journal of Experimental Education*. https://doi.org/10.1080/00220973.2020.1711697

Fiero, M. H., Huang, S., & Bell, M. L. (2016). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*, *17*(72). https://doi.org/10.1186/s13063-016-1201-z

Finch, W. H., & Miller, J. E. (2019). The Use of Incorrect Informative Priors in the Estimation of MIMIC Model Parameters with Small Sample Sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 497–508. https://doi.org/10.1080/10705511.2018.1553111

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Galbraith, S., M.Stat, & Marschner, I. C. (2002). Guidelines for the design of clinical trials with longitudinal outcomes. *Controlled Clinical Trials*, *23*(3), 257–273. https://doi.org/10.1016/S0197-2456(02)00205-2

García-Donato, G., & Chen, M.-H. (2005). Calibrating Bayes Factor Under Prior Predictive Distributions. *Statistica Sinic*, *15*, 359–380.

Garnier-Villarreal, M., & Jorgensen, T. D. (2019). Adapting Fit Indices for Bayesian Structural Equation Modeling: Comparison to Maximum Likelihood. *Psychological Methods*. https://doi.org/10.1037/met0000224

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are Fit Indices Really Fit to Estimate the Number of Factors With Categorical Variables? Some Cautionary Findings via Monte Carlo Simulation. *Psychological Methods*, *21*(1), 93–111. https://doi.org/10.1037/met0000064

Geisser, S., & Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, *74*, 153–160. https://doi.org/10.1080/01621459.1979.10481632

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework. *Psychological Methods*, *19*(1), 79–91. https://doi.org/10.1037/a0032138.supp

Gelfand, A. E., & Dey D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(3), 501–514.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*(3), 445–450. https://doi.org/10.1214/08-BA318

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electronic Journal of Statistics*, *7*(1), 2595–2602.

https://doi.org/10.1214/13-EJS854
Gelman, A., Bois, F., & Jiang, J. (1996). Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions. *Journal of the American Statistical Association*, *91*(436), 1400–1412. https://doi.org/10.1080/01621459.1996.10476708
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman & Hall/CRC.
Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, *6*(4), 733–760.
Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. http://www.jstor.org/stable/2246093
Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 1–13. https://doi.org/10.3390/e19100555
Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.
Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, *4*, 641–649.
Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A Language and Program for Complex Bayesian Modelling. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *43*(1), 169–177.
Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530
Gronau, Q. F., Singmann, H., & Wagenmakers, E. J. (2020). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, *92*(10). https://doi.org/10.18637/jss.v092.i10
Hanson, L., VandeVusse, L., Garnier-Villarreal, M., McCarthy, D., Jerofke-Owen, T., Malloy, E., & Paquette, H. (2020). Validity and Reliability of the Antepartum Gastrointestinal Symptom Assessment Instrument. *Journal of Obstetric, Gynecologic & Neonatal Nursing*. https://doi.org/10.1016/j.jogn.2020.02.006
Harindranath, R. M., & Jacob, J. (2018). Bayesian structural equation modelling tutorial for novice management researchers. *Management Research Review*, *41*(11), 1254–1270. https://doi.org/10.1108/MRR-11-2017-0377
Harring, J. R., Mcneish, D. M., & Hancock, G. R. (2017). Using Phantom Variables in Structural Equation Modeling to Assess Model Sensitivity to External Misspecification. *Psychological Methods*, *22*(4), 616–631. https://doi.org/10.1037/met0000103
Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97.
Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking Misfit in Confirmatory Factor Analysis by Increasing Unique Variances: A Cautionary Note on the Usefulness of Cutoff Values of Fit Indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM Fit Indexes With Respect to Violations of Uncorrelated Errors. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 36–50. https://doi.org/10.1080/10705511.2012.634710

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*(6), 1109–1144.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623. http://mcmc-jags.sourceforge.net

Hoijtink, Herbert, Mulder, J., Van Lissa, C., & Gu, X. (2019). A Tutorial on Testing Hypotheses Using the Bayes Factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A Comparison of ML, WLSMV, and Bayesian Methods for Multilevel Structural Equation Models in Small Samples: A Simulation Study. *Multivariate Behavioral Research*, *51*(5), 661–680. https://doi.org/10.1080/00273171.2016.1208074

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating Model Fit in Bayesian Confirmatory Factor Analysis With Large Samples: Simulation Study Introducing the BRMSEA. *Educational and Psychological Measurement*, *78*(4), 537–568. https://doi.org/10.1177/0013164417709314

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford Press.

Hsu, H.-Y., Kwok, O.-M., Lin, H., & Acosta, S. (2015). Detecting Misspecified Multilevel Structural Equation Models with Common Fit Indices: A Monte Carlo Study. *Multivariate Behavioral Research*, *50*(2), 197–215. https://doi.org/10.1080/00273171.2014.977429

Hu, L.-T., & Bentler, P. M. (1998). Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification. *Psychological Methods*, *3*(4), 424–453.

Jang, G. H. (2010). *Invariant procedures for model checking, checking for prior-data conflict and Bayesian inference.* Ph.D. Thesis, University of Toronto, Toronto, ON, Canada.

Jaynes, E. T. (1968). Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 227–241. https://doi.org/10.1109/TSSC.1968.300117

Jeffreys, H. (1961). *Theory of probability*. Clarendon Press.

Kaplan, D. (1989). Model Modification in Covariance Structure Analysis: Application of the Expected Parameter Change Statistic. *Multivariate Behavioral Research*, 24–3. https://doi.org/10.1207/s15327906mbr2403_2

Kaplan, D. (1990). Evaluating and Modifying Covariance Structure Models: A Review and Recommendation. *Multivariate Behavioral Research*, *25*(2), 137–155. https://doi.org/10.1207/s15327906mbr2502_1

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions*. Sage Publications.

Kaplan, D. (2014). *Bayesian Statistics for the Social Sciences*. Guilford Press.

Kaplan, D., & Depaoli, S. (2012). Bayesian Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). Guilford

Press. https://doi.org/10.1016/B978-044452044-9/50011-2
Kass, R. E., & Greenhouse, J. B. (1989). [Investigating Therapies of Potentially Great Benefit : ECMO]: Comment: A Bayesian Perspective. *Institute of Mathematical Statistics*, *4*(4), 310–317.
Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572
Kenny, D. A., & Mccoach, D. B. (2003). Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling*, *10*(3), 333–351. https://doi.org/10.1207/S15328007SEM1003_1
Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217.
Kline, R. B. (2015). *Principles and practice of structural equation modeling* (Fourth ed.). Guilford Press.
Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.
Lang, K. M., & Little, T. D. (2018). Principled missing data treatments. *Prevention Science*, *19*(3), 284–294. https://doi.org/10.1007/s11121-016-0644-5
Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Memoire de l'Academie Royale Des Sciences*, *4*, 621–656.
Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Wiley.
Lei, P.-W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and Practical Considerations. *Educational Measurement: Issues and Practice*, *26*(3), 33–43. https://doi.org/10.4324/9780203108550
Leimkuhler, B., & Reich, S. (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press.
Leite, W. L., & Stapleton, L. M. (2006). Sensitivity of fit indices to detect misspecifications of growth shape in latent growth modeling. In *Paper presented at the annual meeting of the American Educational Research Association*.
Lek, K., & van de Schoot, R. (2019). How the choice of distance measure influences the detection of prior-data conflict. *Entropy*, *21*(5). https://doi.org/10.3390/e21050446
Levy, R. (2011). Bayesian Data-Model Fit Assessment for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723
Liang, X. (2020). Prior Sensitivity in Bayesian Structural Equation Modeling for Sparse Factor Loading Structures. *Educational and Psychological Measurement*, 1–34. https://doi.org/10.1177/0013164420906449
Liang, X., Kamata, A., & Li, J. (2020). Hierarchical Bayes Approach to Estimate the Treatment Effect for Randomized Controlled Trials. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164420909885
Liang, X., & Luo, Y. (2019). A Comprehensive Comparison of Model Selection Methods for Testing Factorial Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2019.1649983
Lin, L.-C., Huang, P.-H., & Weng, L.-J. (2017). Selecting Path Models in SEM: A Comparison of Model Selection Criteria. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(6), 855–869.

https://doi.org/10.1080/10705511.2017.1363652
Little, R. J. (2021). Missing Data Assumptions. *Annual Review of Statistics and Its Application*, *8*, 89–107. https://doi.org/10.1146/annurev-statistics-040720
Little, R. J. A., & Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Method and Research*, *18*(2), 292–326.
Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (Second Edi). John Wiley & Sons, Inc. www.copyright.com.
Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375. https://doi.org/10.1016/j.jmp.2008.03.002
Liu, C. W., & Wang, W. C. (2017). Non-ignorable missingness item response theory models for choice effects in examinee-selected items. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 499–524. https://doi.org/10.1111/bmsp.12097
Lu, Z.-H., Chow, S.-M., & Loken, E. (2017). A Comparison of Bayesian and Frequentist Model Selection Methods for Factor Analysis Models. *Psychological Methods*, *22*(2), 361–381. https://doi.org/10.1037/met0000145
Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. J. (2012). *The BUGS book : a practical introduction to Bayesian analysis*. Chapman and Hall/CRC.
Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. https://doi.org/10.1002/sim.3680
MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and Cautions in Implementing Bayesian Structural Equation Modeling. *Psychological Methods*, *17*(3), 340–345. https://doi.org/10.1037/a0027131
Mahler, C. (2011). *The effects of misspecification type and nuisance variables on the behaviors of population fit indices used in structural equation modeling* [University of British Columbia]. https://doi.org/10.14288/1.0105120
Marcoulides, K. M. (2018). Careful with Those Priors: A Note on Bayesian Estimation in Two-Parameter Logistic Item Response Theory Models. *Measurement: Interdisciplinary Research and Perspectives*, *16*(2), 92–99. https://doi.org/10.1080/15366367.2018.1437305
Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533–558. https://doi.org/10.1007/s11336-016-9552-7
Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2017). Assessing Fit in Structural Equation Models: A Monte-Carlo Evaluation of RMSEA Versus SRMR Confidence Intervals and Tests of Close Fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 389–402. https://doi.org/10.1080/10705511.2017.1389611
Mcneish, D. M. (2016). Using Data-Dependent Priors to Mitigate Small Sample Bias in Latent Growth Models: A Discussion and Illustration Using Mplus. *Journal of Educational and Behavioral Statistics*, *41*(1), 27–56. https://doi.org/10.3102/1076998615621299
McNeish, D. M. (2016). On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

Mcneish, D. M., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

Mcneish, D. M., & Hancock, G. R. (2018). The Effect of Measurement Quality on Targeted Structural Model Fit Indices: A Comment on Lance, Beck, Fan, and Carter (2016). *Psychological Methods*, *23*(1), 184–190. https://doi.org/10.1037/met0000157

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*(3), 279–300. https://doi.org/10.1207/s15327574ijt0503

Meehl, P. E. (2009). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry* , *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meng, X.-L. (1994). Posterior Predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/http://doi.org/10.18637/jss.v085.i04

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, *42*(5), 869–874. https://doi.org/10.1016/j.paid.2006.09.022

Miočević, M. (2019). A Tutorial in Bayesian Mediation Analysis with Latent Variables. *Methodology*, *15*(4), 137–146. https://doi.org/10.1027/1614-2241/a000177

Miočević, M., Levy, R., Mackinnon, D. P., & Mio Cevi C A, M. (2020). Different Roles of Prior Distributions in the Single Mediator Model with Latent Variables. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2019.1709405

Morey, R. D., & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 406–419. https://doi.org/10.1037/a0024377

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition*. Muthén & Muthén.

Nakadai, R., Nyman, T., Hashimoto, K., Iwasaki, T., & Valtonen, A. (2020). *Climate, species richness, and body size drive geographical variation in resource*. https://doi.org/10.1101/2020.01.09.899922

Neal, R. (2011). MCMC Using Hamiltonian Dynamics. In Steve P Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 116–162). Chapman; Hall/CRC.

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*(1), 143–153.

https://doi.org/10.1177/0165025415618275
Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *46*, 1148–1172. https://doi.org/10.1007/s11747-018-0602-9
Nott, D. J., Xueou, W., Evans, M., & Englert, B.-G. (2016). Checking for prior-data conflict using prior to posterior divergences. *ArXiv*.
O Roberts, G. (1996). Markov Chain Concepts Related to Sampling Algorithms. In W. R. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 45–58). Chapman & Hall.
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
Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Missing Data in Educational Research*, *74*(4), 525–556.
Phipps, D. J., Hagger, M., & Hamilton, K. (2020). Validity and Reliability of the Antepartum Gastrointestinal Symptom Assessment Instrument. *Appetite*, *150*. https://doi.org/10.1016/j.appet.2020.104668
Plummer, M. (2017). *JAGS Version 4.3.0 user manual*. https://doi.org/10.5604/15093492.1135044
Plummer, M. (2019). *rjags: Bayesian graphical models using MCMC. R package version 4-10* (Version 4-8). CRAN.R-project.org/package=rjags
Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. https://doi.org/10.1214/12-BA730
Press, S. J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications* (Second edi). John Wiley & Sons, Inc.
R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/
Raftery, A. E., & Lewis, S. M. (1992). Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo. *Statistical Science*, *7*(4), 493–497. https://doi.org/10.1214/ss/1177011143
Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Harvard University Graduate School of Business Administration (Division of Research); Bailey & Swinfen.
Rioux, C., & Little, T. D. (2019). Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development*. https://doi.org/10.1177/0165025419880609
Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling Omitted and Not-Reached Items in IRT Models. *Psychometrika*, *82*(3), 795–819. https://doi.org/10.1007/s11336-016-9544-7
Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.
Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing Mistakes in Psychological Science. *Advances in Methods and Practices in Psychological Science*, *2*(1), 3–11. https://doi.org/10.1177/2515245918801915

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://academic.oup.com/biomet/article/63/3/581/270932

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or Not to Bayes, From Whether to When: Applications of Bayesian Methodology to Modeling. *Structural Equation Modeling*, *11*(3), 452–483. https://doi.org/10.1207/s15328007sem1103

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Ryu, E., & West, S. G. (2009). Level-Specific Evaluation of Model Fit in Multilevel Structural Equation Modeling. *Structural Equation Modeling*, *16*, 583–601. https://doi.org/10.1080/10705510903203466

Saris, W. E., Satorra, A., & Van Der Veld, W. M. (2009a). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Saris, W. E., Satorra, A., & Van Der Veld, W. M. (2009b). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Savalei, V. (2012). The Relationship Between Root Mean Square Error of Approximation and Model Misspecification in Confirmatory Factor Analysis Models. *Educational and Psychological Measurement*, *72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*(1), 37–52.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Schwerdtfeger, A. R., Rominger, C., & Obser, P. D. (2020). A shy heart may benefit from everyday life social interactions with close others: An ecological momentary assessment trial using Bayesian multilevel modeling. *Biological Psychology*, *152*, 107864.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research*, *58*, 935–943. https://doi.org/10.1016/j.jbusres.2003.10.007

Shi, Dexin, Lee, T., & Maydeu-Olivares, A. (2019). Understanding the Model Size Effect on SEM Fit Indices. *Educational and Psychological Measurement*, *79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Shi, Dexin, Maydeu-Olivares, A., & Distefano, C. (2018). The Relationship Between the Standardized Root Mean Square Residual and Model Misspecification in Factor Analysis Models. *Multivariate Behavioral Research*, *53*(5), 676–694. https://doi.org/10.1080/00273171.2018.1476221

Shi, Dingjing, & Tong, X. (2017). The Impact of Prior Information on Bayesian Latent Basis Growth Model Estimation. *SAGE Open*. https://doi.org/10.1177/2158244017727039

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge

Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216

Smid, S. C., Depaoli, S., & van de Schoot, R. (2019). Predicting a Distal Outcome Variable From a Latent Growth Model: ML versus Bayesian Estimation. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2019.1604140

Smid, S. C., McNeish, D. M., Miočević, M., & van de Schoot, R. (2019). Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2019.1577140

Smid, S. C., & Winter, S. D. (2020). Dangers of the Defaults: A Tutorial on the Impact of Default Priors When Using Bayesian SEM With Small Samples. *Frontiers in Psychology*, *11*(December), 287–290. https://doi.org/10.3389/fpsyg.2020.611963

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(1), 3–23. https://doi.org/10.1111/j.2517-6161.1993.tb01466.x

Song, X. Y., & Lee, S. Y. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology*, *56*(3), 135–148. https://doi.org/10.1016/j.jmp.2012.02.001

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*(4), 583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *76*(3), 485–493. https://doi.org/10.1111/rssb.12062

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 357. https://doi.org/10.2307/2983527

Stan Development Team. (2020). *RStan: the R interface to Stan. R package version 2.21.2*. http://mc-stan.org/

Stan Development Team. (2021). *21. Reproducibility*. Stan Reference Manual. https://mc-stan.org/docs/2_26/reference-manual/reproducibility-chapter.html

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/ s15327906mbr2502_4

Steiger, J. H., & Lind, J. (1980). Statistically-based tests for the number of common factors. *Paper Presented at the Meeting of the Psychometric Society*.

Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2015). The Prowess and Pitfalls of Bayesian Structural Equation Modeling: Important Considerations for Management Research. *Journal of Management*, *41*(2), 491–520. https://doi.org/10.1177/0149206314551962

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. https://doi.org/10.1080/01621459.1987.10478458

Tarka, P. (2018). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality and Quantity*, *52*(1), 313–354. https://doi.org/10.1007/s11135-017-0469-8

Thiele, C. (2021). *cutpointr: Determine and Evaluate Optimal Cutpoints in Binary Classification Tasks.* (R package version 1.1.0.). https://cran.r-project.org/package=cutpointr

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

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., van Loey, N. E., & Lle Zondervan-Zwijnenburg, M. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*(1), 25216. https://doi.org/10.3402/ejpt.v6.25216

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Review Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00001-2

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. G. (2014). A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

van de Schoot, R., Veen, D., Smeets, L., Winter, S. D., & Depaoli, S. (2020). A Tutorial on Using The Wambs Checklist to Avoid The Misuse of Bayesian Statistics. In R. van de Schoot & M. Miočević (Eds.), *Small Sample Size Solutions: A guide for applied researchers and practitioners* (pp. 30–49). Routledge. https://doi.org/10.4324/9780429273872-4

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A Systematic Review of Bayesian Articles in Psychology: The Last 25 Years. *Psychological Methods*, *22*(2), 217–239.

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388. https://doi.org/10.1037/met0000162

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2010.07.003

Veen, D., Egberts, M. R., van Loey, N. E. E., & van de Schoot, R. (2020). Expert Elicitation for Latent Growth Curve Models: The Case of Posttraumatic Stress Symptoms Development in Children With Burn Injuries. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01197

Veen, D., Stoel, D., Schalken, N., Mulder, K., & van de Schoot, R. (2018). Using the Data Agreement Criterion to rank Experts' beliefs. *Entropy*, *20*(8), 1–17. https://doi.org/10.3390/e20080592

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. In *arXiv*. arXiv. https://doi.org/10.1214/20-BA1221

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Lee, M. D., Lodewyckz, T., & Iverson, G. (2008). Bayesian Versus Frequentist Inference. In H Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian Evaluation of Informative Hypotheses* (pp. 181–207). Springer. https://doi.org/10.1007/978-0-387-09612-4_9

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, *211*(1–2), 1–10. https://doi.org/10.1016/j.ecolmodel.2007.10.030

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). Guilford Press.

Wetherill, G. B. (1961). Bayesian Sequential Analysis. *Biometrika*, *48*(3/4), 281–292.

Wu, W., & West, S. G. (2010). Sensitivity of Fit Indices to Misspecification in Growth Curve Models. *Multivariate Behavioral Research*, *45*(3), 420–452. https://doi.org/10.1080/00273171.2010.483378

Wu, W., & West, S. G. (2013). Detecting Misspecification in Mean Structures for Growth Curve Models: Performance of Pseudo R 2 s and Concordance Correlation Coefficients. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 455–478. https://doi.org/10.1080/10705511.2013.797829

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*(3), 183–201. https://doi.org/10.1037/a0015858.supp

Xia, Y., & Yang, Y. (2018). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1055-2

Young, K. D. S., & Pettit, L. I. (1996). Measuring Discordancy between Prior and Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(4), 679–689. https://doi.org/10.1111/j.2517-6161.1996.tb02107.x

Yuan, K.-H., Zhang, Z., & Deng, L. (2019). Fit Indices for Mean Structures With Growth Curve Models. *Psychological Methods*, *24*(1), 36–53. https://doi.org/10.1037/met0000186.supp

Yuan, Y., & Mackinnon, D. P. (2009). Bayesian Mediation Analysis. *Psychological Methods*, *14*(4), 301–322. https://doi.org/10.1037/a0016972

Zeman, J. L., Dallaire, D. H., Folk, J. B., & Thrash, T. M. (2018). Maternal Incarceration, Children's Psychological Adjustment, and the Mediating Role of Emotion Regulation. *Journal of Abnormal Child Psychology*, *46*(2), 223–236. https://doi.org/10.1007/s10802-017-0275-8

Zhang, X., & Savalei, V. (2020). Examining the effect of missing data on RMSEA and CFI under normal theory full-information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 219–239. https://doi.org/10.1080/10705511.2019.1642111

Zhu, X., & Stone, C. A. (2012). Bayesian Comparison of Alternative Graded Response Models for Performance Assessment Applications. *Educational and Psychological Measurement*, *72*(5), 774–799. https://doi.org/10.1177/0013164411434638

Zitzmann, S., & Hecht, M. (2019). Going Beyond Convergence in Bayesian Estimation: Why Precision Matters Too and How to Assess It. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 646–661. https://doi.org/10.1080/10705511.2018.1545232

Zondervan-Zwijnenburg, M., Depaoli, S., Peeters, M., & Van de Schoot, R. (2018). The Performance of Maximum Likelihood and Bayesian Estimation With Small and Unbalanced Samples in a Latent Growth Model. *Methodology*.

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (2017). Where Do Priors Come From? Applying Guidelines to Construct Informative Priors in Small Sample Research. *Research in Human Development ISSN:*, *14*(4), 305–320. https://doi.org/10.1080/15427609.2017.1370966

Appendix A: A Note on the use of Prior-Predictive Samples with Diffuse Priors for SEM

Researchers should be aware that prior-predictive samples generated through the package 'rstan' for SEMs with diffuse priors on all parameters may be affected by factors not directly related to the observed data, priors, or specified model. Specifically, the developers of Stan (Stan Development Team, 2021) state that Stan results will only be exactly reproducible if all of the following components are identical:

1. Stan version
2. Stan interface (RStan, PyStan, CmdStan) and version, plus version of interface language (R, Python, shell)
3. Versions of included libraries (Boost and Eigen)
4. Operating system version
5. Computer hardware including CPU, motherboard, and memory
6. C++ compiler, including version, compiler flags, and linked libraries
7. Same configuration of call to Stan, including random seed, chain ID, initialization, and data

While researchers can control some of these components directly (e.g., Stan version and random seed), other components are more difficult to control and may change unexpectedly as time passes. For example, if analyses are run on a high-performance cluster (HPC) managed by external parties (e.g., the university), one may not be able to control the operating system version or hardware used.

For this dissertation, the issue with reproducibility arose as I was compiling the results for Study 2. For efficiency, I used several computers (e.g., a simulation computer running Windows 10 and the UC Merced HPC running Linux) to run the simulation cells for Study 2. As I was comparing the results across cells, I noticed clear discrepancies for the prior-predictive samples generated for the fully diffuse prior specification. Further, small discrepancies existed for conditions in which a diffuse prior for one parameter was combined with an informative prior (aligned or diverging) for the other parameter. I systematically adjusted different aspects of my code to determine what the cause of these discrepancies was. From this examination, I could draw several conclusions:

1. Differences between computers did not affect the posterior samples used to compute the posterior estimate bias, RMSE, DAC, or BF. Thus, the reproducibility challenges existed only for the prior-predictive samples used to compute the prior-predictive *p*-values.
2. Prior specifications with two informative priors (aligned or diverging) were not affected by differences between computers.
3. For a specific computer, the versions of R, Stan, and the package 'rstan' did not affect the prior-predictive samples generated with diffuse prior specifications. However, the random seed used in the sampling() function did affect prior-predictive samples generated with the fully diffuse prior specification. The impact of the random seed was less apparent for prior specifications with one informative prior.

4. Ensuring that the same versions of R, Stan, and the package ‘rstan’ (in addition to the same random seed, chain ID, and initialization) were used across computers did not resolve the reproducibility issues for prior-predictive samples generated with diffuse prior specifications.

Thus, researchers should keep in mind that any conclusions regarding the appropriateness of (fully or partially) diffuse prior specifications may depend on the specific computer (and to a lesser extent the random seed specified in the sampling() function) they used to generate the prior-predictive samples. For example, in Study 2, the prior-predictive *p*-value of the mean of *y1* for the diffuse prior on the mean intercept and aligned moderately informative prior on the mean slope specification with $n = 500$ indicated that there was prior-data disagreement for 100% of replications on one computer system (as reported in the Results section), but for 0% of replications on a second computer system.

Although diffuse priors may not reflect prior-data disagreement, examining their appropriateness may still be necessary as diffuse priors can sometimes affect posterior estimates in unexpected ways (Depaoli, 2013; Depaoli & Clifton, 2015; Smid & Winter, 2020; van Erp et al., 2018). Therefore, prior-predictive samples generated for models with (fully or partially) diffuse priors should be interpreted with caution. Furthermore, researchers should provide specific details regarding their random seeds and computer setup when they present their findings so that results can be replicated, or at least so that the reader knows the settings implemented for the specific study. All code (including random seeds) used to generate the prior-predictive samples reported in Study 2 can be found on the OSF page connected to this dissertation. Prior-predictive samples were generated with R version 4.0.3, ‘rstan’ package version 2.21.2, on the UC Merced HPC, which runs CentOS Linux release 7.8.2003 (Core) with kernel 3.10.0-1127.19.1.el7.x86_64. The UC Merced HPC is supported by the National Science Foundation (Grant No. ACI-1429783).