Bayesian SEM with Small Samples

Precautions and Guidelines

Sanne C. Smid

Bayesian SEM with Small Samples

Precautions and Guidelines

Sanne C. Smid

Bayesian SEM with Small Samples: Precautions and Guidelines
Proefschrift Universiteit Utrecht - Met samenvatting in het Nederlands.

DOI: https://doi.org/10.33540/1755
ISBN: 978-94-6483-080-4
Printing and cover DTP by: Ridderprint
Cover design by: Sanne C. Smid

The watercolor on the cover shows an abstract representation of a Bayesian analysis, where observed data (blue color) is combined with prior information (fuchsia color).

Bayesian SEM with Small Samples

Precautions and Guidelines

Bayesiaanse Structurele Vergelijkingsmodellen
met Kleine Steekproeven
Voorzorgsmaatregelen en Richtlijnen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 2 juni 2023 des ochtends te 10.15 uur

door

Sanne Christina Smid

geboren op 7 november 1989
te Utrecht

Promotoren: Prof. dr. A.G.J. van de Schoot
Prof. dr. L.D.N.V. Wijngaards-de Meij

Beoordelingscommissie: Prof. dr. A.L. van Baar
Prof. dr. S. van Buuren
Dr. S.J. van Erp
Prof. dr. I.G. Klugkist
Prof. dr. E.J. Wagenmakers

The studies in this thesis were funded by the Netherlands Organization for Scientific Research (project NWO-VIDI-452-14-006).

There is nothing to do, just be
Fia

Table of Contents

Introduction

Nowadays, it seems there is more than enough data available about everything. However, even in times of *big data*, there are situations where it is challenging to collect *enough* data. Think of naturally small populations, such as people with rare diseases or young children with severe burn injuries (see e.g., Mitani & Haneuse, 2020; Veen, Egberts, Van Loey, & Van de Schoot, 2020). Or hard to access target groups, such as people with addiction problems or low literacy, survivors of violence, or undocumented migrants (see e.g., Bonevski et al., 2014; Vollebregt, Scholte, Hoogerbrugge, Bolhuis, & Vermeulen, 2022). Small samples can also be due to financial constraints, think of studies in which expensive MRI scans are used and collecting enough data is simply too expensive (see e.g., Turner, Paul, Miller, & Barbey, 2018). Small samples are inevitable in situations like this.

When a sample is considered small, is discussed in Chapters 1 and 4 of this dissertation. For now, it is important to know that small samples can cause big problems. Statistical methods require a certain amount of data to perform well. This dissertation focuses on Structural Equation Models (SEMs) - a flexible modeling framework in which latent variables (i.e., variables that are not directly observed, such as quality of life or happiness) can be modeled. An example of such a model is a latent growth model, in which the development of a latent variable can be investigated over time. Running an SEM without enough data can result in extremely low levels of power, meaning that effects in the data are probably not detected (see e.g., Cohen, 1988). Other problems that can occur are inaccurate parameter estimates, inadmissible parameter estimates, or the absence of parameter estimates due to non-convergence of the model (see e.g., Boomsma, 1985; Nevitt & Hancock, 2004).

One way to avoid these small-sample-problems is to simplify the research question and statistical model, and for instance only report descriptive statistics or use a very simple statistical test. However, this is not desirable as complex and essential research questions about naturally small populations and hard to access target groups cannot be answered in that way, and important information is missed out on.

Another option to circumvent small-sample-problems that does not involve simplifying the model or research question, could be the use of Bayesian methods. In the Bayesian framework, observed data is combined with *prior knowledge*. Prior knowledge is information about the model of interest and corresponding parameters that exist *before* the data is collected. This knowledge can for example be based on previous studies, opinions of experts in the field, or on information about the parameters based on the scale that will be used to collect data (see e.g., Lek & Van de Schoot, 2018; Zondervan-Zwijnenburg, Peeters, Depaoli, & Van de Schoot, 2017). This information is captured in a distribution, the so-called prior distribution. The combination of prior distributions and the observed data is what we call the posterior, this is the result of a Bayesian analysis. With a small sample size, priors have a relatively larger impact on the posterior than with a large sample size. For a basic introduction to Bayesian statistics, we refer to Gelman et al. (2014) and Kruschke (2015). For an introduction to Bayesian SEM, we refer to Kaplan & Depaoli (2012), Asparouhov & Muthén (2010), and Depaoli (2021).

The inclusion of prior knowledge can increase the amount of information that is available in the analysis, simply because more information is added through the specification of prior distributions. Also, Bayesian methods do not rely on large sample techniques in contrast to the classical frequentist methods (see e.g., Gelman et al., 2014). Therefore, Bayesian estimation is *in theory* more suitable for small samples. Recently, more and more often researchers switch to a Bayesian approach to *deal* with their small sample sizes (McNeish, 2016a; Van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). However, this switch is not without problems, and that is where the studies of this dissertation come in.

Outline of this Dissertation

In the four studies in this dissertation, we create an overview of the performance of Bayesian Structural Equation Models (SEMs) with small samples compared to classical frequentist methods. In addition, we discuss precautions and provide guidelines on how to use Bayesian SEM in a *thoughtful* way when samples are small.

All chapters have their own page at the Open Science Framework (OSF). Here, supplemental files can be found as well as scripts containing annotated R or M*plus* code to reproduce the results. The link to the corresponding OSF page is provided at the beginning of each chapter.

In **Chapter 1**, the results are presented from an extensive systematic literature review on the performance of Bayesian and frequentist estimation methods under small samples for SEMs. The results of this systematic review are widely applicable, as in the included studies a variety of SEMs was investigated. We present an overview of the included studies, as well as the models of interest, which sample sizes are considered to be small according to the authors of the included studies, and aggregate the information from the included studies. We end the chapter with recommendations for researchers on analyzing a small sample size and on how to specify thoughtful prior distributions. The study described in this chapter is published in Structural Equation Modeling: A Multidisciplinary Journal. Note that as a response on this publication, a comment paper was written by Zitzmann, Lüdtke, Robitzsch and Hecht (2021). Also, the data set containing all screened references is openly available at the OSF, and can be used for other purposes as well (see e.g., the simulation study by Ferdinands, 2021).

In the systematic literature review, we did not come across any simulation studies investigating latent growth models with a long-term outcome. In this model, growth processes can be modeled over time (e.g., the development of posttraumatic stress symptoms over time), and it can be investigated whether the initial starting point (i.e., posttraumatic stress score on time point 1) or growth process (i.e., the development of posttraumatic symptoms over time) can be indicators for another variable measured later in time (e.g., quality of life). This type of model allows researchers to assess longer-term patterns, and to detect the need to start a (preventive) treatment or intervention in an early

stage. In **Chapter 2**, we discuss the results of a simulation study in which we investigate the performance of Bayesian and frequentist estimation for a latent growth model with a long-term (so-called 'distal') outcome variable. The research described in this chapter is published in Structural Equation Modeling: A Multidisciplinary Journal.

In **Chapter 3**, we touch on promising estimation methods for small samples within the frequentist framework: twostep modeling and factor regression score. Both methods are available in the software package `lavaan` in R (Rosseel, 2012).[1] In a simulation study, we investigate the performance of twostep modeling and factor regression score, and compare them to Maximum Likelihood estimation and Bayesian estimation. This study is published as a book chapter in the book 'Small Sample Size Solutions: A Guide for Applied Researchers' (2020), edited by van de Schoot and Miočeviç.

In **Chapter 4**, we present the *dangers of the defaults*: a non-technical tutorial in which we discuss the risks of using Bayesian estimation while blindly relying on built-in software default priors when samples are small. Also, we demonstrate an online educational Shiny app (Smid & Winter, 2020, available via https://osf.io/m6byv/), in which users can play around with varying sample sizes and prior settings to investigate the impact of priors on the results. Finally, we present guidelines on how to recognize 'misbehaving' and 'behaving' priors after the Bayesian analysis is conducted. This study is published in a special issue on Bayesian methods in psychology in the journal Frontiers in Psychology.

[1] Twostep modeling and FSR are both variants of the Structural-after-Measurement (SAM) approach in the software package `lavaan`. In a nutshell, 'twostep' is gobal SAM, and 'fsr' is local SAM. For more information about SAM, we refer to Rosseel & Loh (2022).

Chapter 1

Bayesian vs Frequentist Estimation for SEMs with Small Samples: A Systematic Review

This chapter is published as Smid, S. C., McNeish, D., Miočeviç, M., & van de Schoot, R. (2020). Bayesian Versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131-161. https://doi.org/10.1080/10705511.2019.1577140

Author Contributions: RvdS and SS designed the study. SS carried out the largest part of the screening of abstracts and full-texts. All doubts were discussed with DM, MM and/or RvdS. SS carried out the qualitative synthesis, and wrote and revised the manuscript with feedback and input of DM, MM and RvdS. RvdS supervised the project.

Online Data Archive and Supplementary Files: https://osf.io/7mght/

Abstract

In small sample contexts, Bayesian estimation is often suggested as a viable alternative to frequentist estimation, such as maximum likelihood estimation. Our systematic literature review is the first study aggregating information from numerous simulation studies to present an overview of the performance of Bayesian and frequentist estimation for structural equation models with small sample sizes. We conclude that with small samples, the use of Bayesian estimation with diffuse default priors can result in severely biased estimates - the levels of bias are often even higher than when frequentist methods are used. This bias can only be decreased by incorporating prior information. We therefore recommend against naively using Bayesian estimation when samples are small, and encourage researchers to make well-considered decisions about all priors. For this purpose, we provide recommendations on how to construct thoughtful priors.

1.1 Introduction

The use of Bayesian estimation is on the rise in many scientific fields (König & Van de Schoot, 2017; Kruschke, Aguinis, & Joo, 2012; Rietbergen, Debray, Klugkist, Janssen, & Moons, 2017; Rupp, Dey, & Zumbo, 2004; Van de Schoot, Winter, et al., 2017), and during the last few decades there has been a "steep increase" in the number of "theoretical, simulation and application papers implementing Bayesian SEM [Structural Equation Modeling]" in psychology (Van de Schoot, Winter, et al., 2017, p. 231). The rise in both applications and methodological studies of Bayesian estimation might be due to the availability in popular software packages and some advantages that Bayesian estimation possesses over its frequentist counterpart, such as the flexibility to include model uncertainty, and to estimate models that are too complex or too computationally demanding for frequentist estimation (see e.g., Kaplan, 2014, pp. 297–290; Van de Schoot, Winter, et al., 2017; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008).[1]

Another popular reason to choose Bayesian estimation is that, unlike frequentist methods (e.g., maximum likelihood (ML) estimation), it does not rely on asymptotic theory (see e.g., Gelman et al., 2014, pp. 83–97; Kaplan, 2014, pp. 285–286). It is often shown that in the context of SEM for small sample sizes, in relation to the complexity of the model, frequentist estimation often results in nonconvergence, inadmissible parameter solutions, and inaccurate estimates. All of these issues might be circumvented by using Bayesian estimation (see e.g., Muthén & Asparouhov, 2012; Wagenmakers et al., 2008). This is a welcoming feature of Bayesian estimation, especially in the social sciences where it can be challenging to collect enough data due to naturally small populations (e.g., Egberts et al., 2016), hard to access target groups (e.g., Coleman et al., 2002), or financial constraints may exist (e.g., Van Lier et al., 2017).[2]

[1] In the current chapter, we assume basic knowledge on Bayesian statistics. For a discussion of the differences between Bayesian and frequentist estimation, see, for example, the chapter on Bayesian and frequentist statistical schools in Kaplan (2014), pp. 285–296. Readers interested in Bayesian statistics are referred to, among many others: Gelman et al. (2014), Kaplan (2014), Kaplan & Depaoli (2013), Kruschke (2015), Lynch (2007), Lee & Wagenmakers (2014), and for recent methodological articles to the two special issues on Bayesian Data Analysis from Psychological Methods (Chow & Hoijtink, 2017; Hoijtink & Chow, 2017).

[2] Although not further discussed in the current study, note that there are several other

Recommendations to use Bayesian over frequentist estimation in small sample contexts are common in the literature. For example, Rupp et al. (2004) mentioned that "Bayesian parameter estimation is more appropriate than ML estimation for smaller sample sizes, because the former do not rely on asymptotic results that are typically not satisfied with psychometric data except in large-scale settings." (p. 446). Kruschke et al. (2012) advised that "Bayesian methods can be used regardless of the overall sample size or relative sample sizes across conditions or groups." (p. 743). Such statements can create the impression that using Bayesian estimation universally solves small sample problems. Although several textbooks on Bayesian estimation stress the important role of prior distributions when Bayesian estimation is used with small samples (e.g., Gelman et al., 2014, p. 88; Kaplan, 2014, p. 291; McElreath, 2016, p. 31), in practice prior distributions are often not carefully chosen, and most empirical researchers rely on default software settings (see e.g., König & Van de Schoot, 2017; McNeish, 2016a; Van de Schoot, Schalken, & Olff, 2017; Van de Schoot, Winter, et al., 2017). Popular software programs, such as: M*plus* (Muthén & Muthén, 1998-2017); SPSS (IBM Corp., 2017); JASP (JASP team, 2018); or the R package `blavaan` (Merkle & Rosseel, 2018), offer Bayesian estimation with diffuse default prior distributions. This permits a *naive* use of Bayesian estimation, which entails that software defaults (e.g., M*plus* default priors) *or* generic rules-of-thumb (e.g., the Inverse Gamma (0.01, 0.01) for variance parameters in multilevel models) are used to specify prior distributions. Naive priors should not be confused with noninformative priors. Some diffuse default priors can act as very informative priors when the sample size is small (see e.g., Gelman, 2006; McNeish, 2016a). In contrast, *thoughtful priors* incorporate previous beliefs about parameters and are adjusted to the specific research situation. These prior distributions could be based on previous studies, meta-analyses or expert opinions and are applicable only to a specific study. In a *thoughtful* way of using Bayes, flat or software default priors can also be used, as long as arguments are provided why this is a suitable prior for this specific parameter, that is, a *thoughtful* choice is made about the prior distributions. A last category are priors based on the data itself, so-called *data dependent priors*. With *data dependent priors*, the model is first fit with a frequentist

techniques to handle small sample sizes in SEM, such as ridge SEM (Yuan, Wu & Bentler, 2011), and three-step estimation (Bakk, Oberski, & Vermunt, 2014).

method (e.g., ML). The estimates of the frequentist estimation are then used as hyperparameters for the prior distributions, often in combination with very large variances to represent the uncertainty about the prior distribution (see e.g., Darnieder, 2011).[3]

1.1.1 Goals of the Study

In the last decade, many simulations studies have investigated the performance of Bayesian estimation for SEM in small samples and compared its performance to frequentist estimation methods. The goal of our systematic review is twofold. The first goal is to provide a comprehensive overview of the performance of Bayesian estimation for SEMs with small samples in comparison to frequentist estimation. Therefore, we report details about the conditions investigated in the included simulation studies, which sample sizes were defined as small by the authors of the studies, and which prior distributions were used. In addition, we aggregate information about coverage, power, and relative bias from all cells across the included simulation studies. Second, we provide recommendations for researchers regarding analyzing small data sets and how to specify *thoughtful* priors.

1.1.2 Organization of the Chapter

The remainder of the chapter is structured as follows: first, the methods used to conduct the systematic review are described, followed by a description of the included studies and the general performance of the investigated estimation methods for SEM with small samples. In addition, we collected and graphically present all the reported coverage, power and relative bias estimates for all parameters from all cells as reported in the included studies. We end with conclusions, a discussion of limitations, and recommendations.

[3]Hyperparameters are the parameters of the prior distribution, for example, the mean and variance in a normal distribution.

1.2 Methods

1.2.1 Inclusion and Exclusion Criteria

We included papers in which a simulation study was used to investigate and compare the performance of Bayesian estimation to frequentist methods in structural equation models with a small sample size. We only included peer-reviewed papers in the field of social sciences. Non-English references were excluded, as well as books, book chapters, conference talks and software manuals. We used the following definitions of the inclusion criteria:

- *Simulation study.* Multiple replicated datasets were analyzed, and results were summarized for all simulated data sets.
- *Bayesian estimation was compared to frequentist estimation methods.* The performance of Bayesian and frequentist estimation was investigated for the exact same model, so that the results can be compared across the two estimation methods.
- *Structural equation models.* Models of interest fall under the umbrella of structural equation models including mediation, confirmatory factor analysis, latent growth, multilevel, and mixture models. Network analysis, machine learning, meta-analysis and item response theory were excluded.
- *Small sample size.* The original authors stated that at least one of the sample sizes in the simulation study represent a small sample size for their specific model.[4] Small sample conditions must have been reported explicitly; aggregated results including small sample conditions were excluded.

1.2.2 Search Strategy

Three approaches containing six searches, were conducted to identify possibly relevant papers, as displayed in Figure 1.1. As the first approach, we used the simulation study papers on small samples which were identified by

[4] An exception is made for two studies in which the authors did not mention that a small sample size was used in the simulation study, while an obviously small sample size was used: 6 and 12 clusters in a multilevel model (Browne & Draper, 2000, 2006).

the systematic review of Van de Schoot, Winter, et al. (2017) on the use of Bayesian estimation in psychology. As a second approach, we sent messages to all subscribers of the mailings lists SEMNET and JISC Multilevel of Listserv 16.0, and posted a message on the online platform ResearchGate. The abstracts of the papers identified from these two approaches (Searches 1-5, see Figure 1.1) were screened and when these met the inclusion criteria, the full-text version was examined. When the inclusion criteria were still met, the paper was included in the qualitative synthesis and in the next search phase, in which the references from the paper were examined as were papers that cited the included paper. Scopus was used to identify the references of the relevant papers, as well as the papers that cited the relevant papers (when the paper was not available in Scopus, Google Scholar was used). These steps were repeated until no new papers were identified. For the first three searches, references that did not meet all our inclusion criteria but did meet the criteria about simulation studies, Bayesian estimation and small samples, were included in the upcoming searches because these papers could still identify relevant references and citations. As a third approach, a final search (Search 6) was carried out using Scopus to identify relevant studies that were published after 2014, because the study of Van de Schoot, Winter, et al. (2017), which was used as the first approach, included studies published until 2015. The exact search strings can be found in the Supplemental File S1 (all Supplementary files are available on the Open Science Framework: https://osf.io/7mght/). The abstracts, followed by the relevant full-texts of the identified records, were screened using the aforementioned inclusion and exclusion criteria.

The first author carried out the screening and as a quality check, a random sample of 10% of abstracts and 20% of full-texts were reviewed by each of the three co-authors, which resulted in very few discrepancies. Disagreements were discussed until the authors agreed. In the end, no additional studies were included in the systematic review after discussion. In Figure 1.2, a summary of the flow charts can be found following Preferred Reporting Items for Systematic reviews and Meta-Analyses [PRISMA; Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group (2009)]. More details of the search are provided online (Supplemental File S1) as well as all identified references and the reason for exclusion (Supplemental File S2). Additionally, separate flowcharts for Searches 1 to 6 are available in Supplemental File S3.

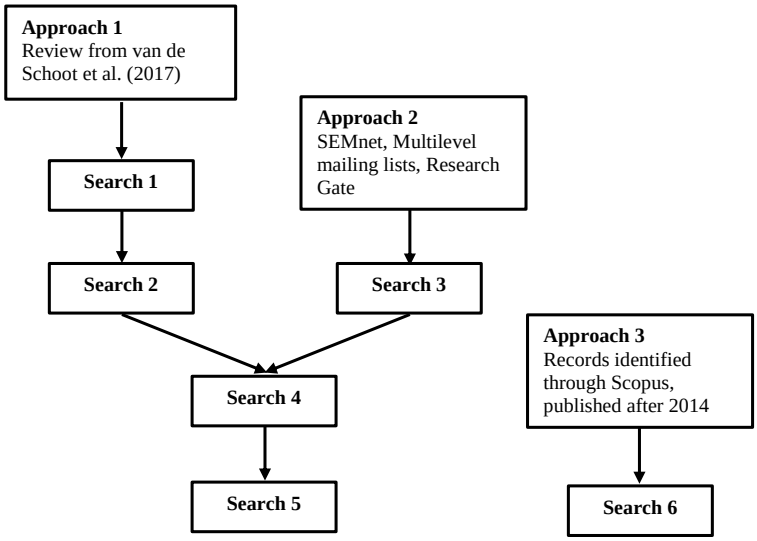


Figure 1.1: The three approaches and six subsequent literature searches to identify relevant references.

1.2.3 Results Search Strategy

A total of 32 studies, described in 27 papers and written by 24 unique groups of authors, met all inclusion criteria and were included in the qualitative synthesis. The following SEMs were investigated in these studies: mediation model ($n = 6$), confirmatory factor analysis ($n = 3$), latent growth model ($n = 6$), multilevel model ($n = 12$), autoregressive model ($n = 1$), and mixture model ($n = 4$). Characteristics of the 32 included studies can be found in Table 1.1. In addition, we collected coverage, power and relative bias for all reported parameters for all cells as reported in the studies.[5] We graphically present these data in Figures 1.3, 1.4 and 1.5.

[5]We used all available results reported in tables in the included papers and appendices. When figures with coverage, power and/or relative bias results were shown in the paper, we contacted the authors to share their simulation results with us. For more details, see Supplemental Table S4.

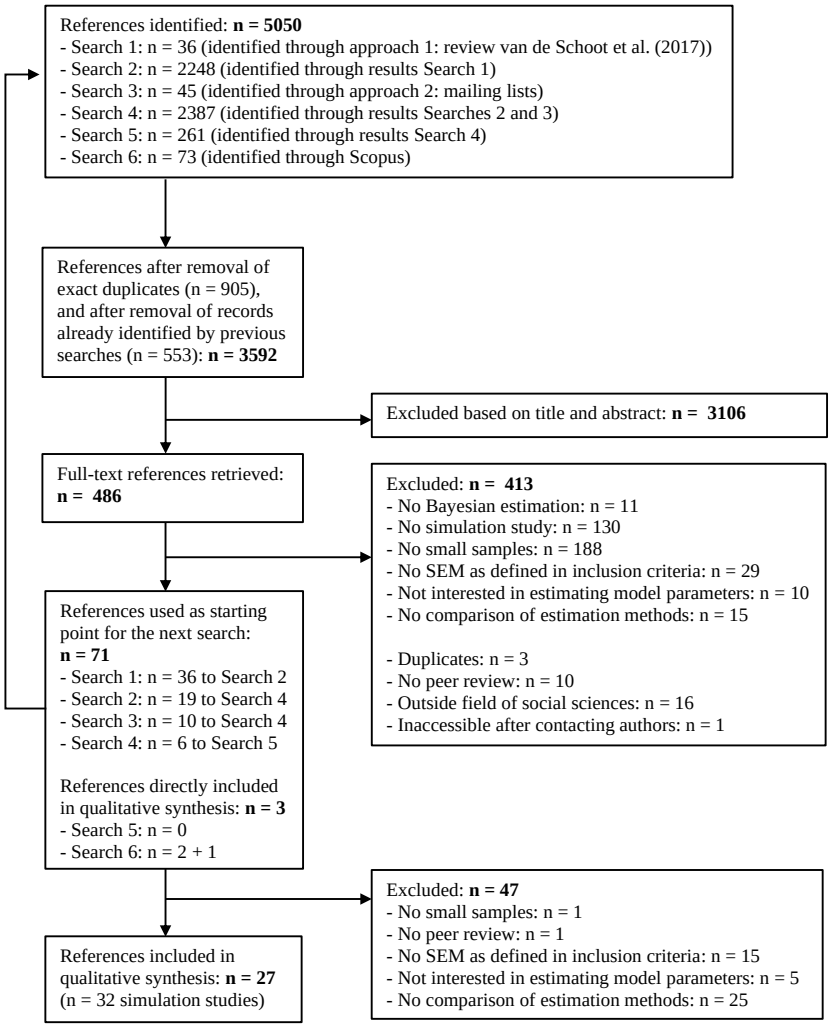


Figure 1.2: Summary flow chart of the search process (based on the PRISMA guidelines). For a detailed description of the exclusion criteria, we refer to the 'Inclusion and exclusion criteria' Section. See Supplemental file S1 for more details about the search process.

Table 1.1: Selected characteristics of simulation studies investigating frequentist and Bayesian estimation methods for SEM with small samples

Study	Model of interest	Estimation methods	Software	Sample Size: Number of Persons/ Clusters	Sample Size: Time Points/ Cluster Size
Mediation Models					
1. Chen et al., 2014*	Mediation model with 3 manifest variables	ML, BayesN	OpenBUGS, M*plus*	**25, 50,** 200	-
2. Chen et al., 2014*	Mediation model with 3 latent variables and continuous indicators	ML, BayesN	OpenBUGS, M*plus*	**50, 100,** 400	-
3. Chen at al., 2015	Mediated-effect model with 3 latent variables and ordinal indicators	RWLS, BayesN, BayesT	M*plus*, OpenBUGS	**100, 200,** 400	-
4. Koopman et al. 2015	Mediation model with 3 manifest variables	OLS, BayesN	MASS, boot, MCMCpack in R	**20, 40, 60, 80**, 100	-
5. Miočević et al. 2017	Single mediator model with 3 manifest variables	OLS, BayesN, BayesT	SAS 9.4, RMediation, SAS PROC MCMC	**20, 40, 60**, 100, 200	-
6. Yuan & MacKinnon, 2009	Mediation model with 3 manifest variables	ML, BayesN, BayesT	WinBUGS	**25, 50,** 100, 200, 1000	-
CFA Models					
7. Natesan, 2015	Ordinal CFA model with 2 factors	RML, WLS, RDWLS, RULS, BayesT	JAGS, LISREL	**42, 63, 84, 105,** 210, 315	-
8. Lee & Song, 2004	Model with 2 overlapping correlated factors; Model with 3 overlapping correlated factors	ML, BayesD	LISREL, BUGS	**32, 48, 64, 80; 44, 66, 80, 110**	-
9. Van Erp et al., 2018	Model with 3 latent variables and mediation effect	ML, BayesN, BayesD	M*plus* 7.2	**35, 75,** 150, 500	-
Latent Growth Models					
10. McNeish, 2016a*	Latent growth model with 2 binary time-invariant exogenous predictors (I, LS)	FIML, REML KR, BayesN, BayesT	M*plus*, SAS PROC MIXED	**20, 30, 50**	4
11. McNeish, 2016b*	Latent basis model and second order growth model (I, LS)	FIML, BayesN, BayesD	M*plus* 7.1	**20, 30, 50**	4
12. McNeish, 2016b*	Latent growth model with 2 binary individual-level predictors (I, LS)	FIML, BayesN	M*plus* 7.1	**20, 30, 50**	4

Study	Model of interest	Estimation methods	Software	Sample Size: Number of Persons/ Clusters	Sample Size: Time Points/ Cluster Size
13. Van de Schoot et al., 2015	Latent growth model including covariate to predict the linear slope (I, LS, QS)	ML, BayesN, BayesT	M*plus* 7.1	**8, 14, 22**	3
14. Zondervan-Zwijnenburg et al., 2019*	Multigroup latent growth model (I, LS, QS)	MLR, BayesT	M*plus* 7.11	Group 1 = **5, 10, 25, 50**; Group 2 = 50, 100, 200, 500, 1000, 2000, 5000, 10.000	4
15. Zondervan-Zwijnenburg et al., 2019*	Multigroup latent growth model (I, LS, QS)	MLR, BayesT	M*plus* 7.11	Group 1 = **50**; Group 2 = 50, 100, 200, 500, 1000, 2000, 5000, 10.000	4
Multilevel Models					
16. Baldwin & Fellingham, 2013	Two-level partially clustered design	REML KR, BayesT	SAS PROC MIXED/ MCMC	**8, 16**	**5, 15**
17. Browne & Draper, 2000	Two-level random-slopes regression model	IGLS, RIGLS, BayesN, BayesD	MLwiN, BUGS	12, 48	(un)balanced, mean = 18
18. Browne & Draper, 2006	Two-level variance-components model	ML, REML, BayesN	MLwiN, WinBUGS	6, 12, 24, 48	(un)balanced, mean = 18
19. Depaoli & Clifton, 2015	Two-level latent covariate model with dichotomous and continuous indicators	MLR /WLSM, BayesN, BayesT	M*plus*	**40**, 50, 100, 200	**5**, 10, 20
20. Farrell & Ludwig, 2008	Two-level response time model	ML, BayesT	N.A.	(i) 20; (ii) 5; (iii) 80	(i) **20**, 80, 500; (ii) 500; (iii) **20**
21. Holtmann et al., 2016	Two-level CFA model with two correlated factors at both levels, continuous and categorical indicators	MLR/ WLSMV, BayesN, BayesT	M*plus* 7 and Mplusautomation in R 3.0.2.	**50**, 100, 150, 200	**2**, 4, 6
22. Hox et al., 2012	Two-level model with one factor and one exogenous predictor	ML (results from other study), BayesN	M*plus* 6.1	**10, 15, 20**	1755
23. Hox et al., 2014	Two-level mediation model	ML, BayesN	M*plus* 7.0	**5, 10, 25**, 50	5, 10
24. McNeish, 2016a*	Two-level model with treatment effect measured at level 2	FIML, REML KR, BayesN, BayesT	M*plus*, SAS PROC MIXED	**8, 10, 14**	7-14

				Sample Size	
Study	Model of interest	Estimation methods	Software	Number of Persons/ Clusters	Time Points/ Cluster Size
25. McNeish & Stapleton, 2016	Two-level model	ML, REML, REML KR, BayesN	SAS PROC MCMC/ MIXED/ GLIMMIX	**4, 8, 10, 14**	7-14, 17-34
26. Stegmueller, 2013	Linear and nonlinear two-level random-intercept models	ML, BayesN	N.A.	**5, 10, 15, 20, 25, 30**	500
27. Tsai & Hsiao, 2008	Two-level model	REML, BayesN	R, SAS PROC GLIMMIX	**15**	**6**
AR Model					
28. Price, 2012	Multivariate autoregressive lag-1 model	MLR, BayesT	M*plus* 6.2	**1, 3, 5, 10, 15**	25, 50, 75, 100, 125
Mixture Models					
29. Depaoli, 2012*	Two-factor model with 2 classes, class separation at measurement level	ML, BayesT	M*plus*	100 (smallest class is **20**), 300, 800	-
30. Depaoli, 2012*	Two-factor model with 2 classes, class separation at structural level	ML, BayesT	M*plus*	100 (smallest class is **20**), 300, 800	-
31. Depaoli, 2013	Growth mixture model with 3 classes (I, LS and in 1 condition also QS)	ML, BayesN, BayesT, BayesD	M*plus* 7	150 (smallest class is **15**), 800	4
32. Serang et al., 2015	Exponential growth mixture model with 2 classes	ML, BayesT	R, OpenBUGS, M*plus* 6.12	**200** (smallest class is **40**), 500, 1000	5, 7, 9

Note. Every line in the table represents one simulation study. * = multiple simulation studies from this paper are included in the qualitative synthesis. - = not applicable, I = intercept, LS = linear slope, QS = quadratic slope, ES = exponential slope. **Bold** = defined as a small sample size by the original authors. Underlined = not defined by original authors, defined by current authors as an obviously small sample size. Bayesian estimation methods abbreviations: BayesN = Bayesian methods with *naive* priors, BayesT = Bayesian methods with *thoughtful* priors, BayesD = Bayesian methods with *data dependent* priors, Frequentist estimation methods abbreviations (in alphabetical order): FIML = Full Information Maximum Likelihood, IGLS = Iterative Generalized Least Squares, ML = Maximum Likelihood, REML = Restricted Maximum Likelihood, REML KR = Restricted Maximum Likelihood with Kenward-Roger correction, RDWLS = Robust Diagonally Weighted Least Squares, RIGLS = Restricted Iterative Generalized Least Squares, RML/ MLR = Robust Maximum Likelihood, RULS = Robust Unweighted Least Squares, RWLS = Robust Weighted Least Squares, WLSM = Weighted Least Squares using a diagonal weight matrix. N.A. in Software column = information on the software program used is not available in the article.

1.2.4 Bayesian vs. Frequentist Methods in Included Studies

In the current study, we distinguish between three types of frequentist estimation methods and three types of prior settings for Bayesian estimation. For the frequentist estimation methods, we differentiate between maximum likelihood (ML), restricted maximum likelihood (REML) and least squares (LS). The ML category subsumes robust ML and full information ML. In the REML category, REML with and without Kenward-Roger correction are included (for more information, see Kenward & Roger, 1997, 2009; McNeish, 2016b). Note that REML with Kenward-Roger correction is often referred to as a "small sample correction" (see e.g., McNeish & Stapleton, 2016, p. 4). Finally, robust weighted least squares or unweighted least squares, all comprise the LS category.

Furthermore, a distinction is made between three types of prior settings for Bayesian estimation. We use the terms: naive (BayesN), thoughtful (BayesT) and data dependent (BayesD) priors. In the current study, the prior setting is categorized as BayesT when information is included in at least one prior distribution. We do not intend to imply that studies using BayesN of BayesD are necessarily lacking though as these approaches are justifiable under some circumstances. Rather, this set of terminology is intended to imply that additional thought was required to specify custom prior distributions instead of relying on defaults, generalized suggestions, or the data to create priors. In Supplemental File S6, the specified prior distributions from all included simulation studies are presented.

Note that within the three Bayesian categories, still different levels of informativeness can occur, as well as different combinations of *naive*, *thoughtful* and *data dependent* priors. However, the study would not benefit from creating subcategories in which only the exact same level of informativeness and combinations of priors occur, as almost each study would end up in a category on its own. Our view is that the three categories we selected are specific enough to discriminate between different types of prior distributions while also allowing for broad conclusions to be readily interpretable.

In the next section, we describe how Bayesian estimation (BayesN, BayesT,

BayesD) performed in comparison to frequentist estimation (ML, REML, LS) in the included studies. We realized that the results in terms of performance of estimation methods, were generally independent of the model. Therefore, we discuss the results across all models together and focus on model specific exceptions. Supplemental Table S6 shows which studies compared which permutations of methods (e.g., which studies compared BayesT to frequentist estimation), and Supplemental Tables S7-S10 include the raw conclusions regarding the performance of the methods in each of the studies.

1.3 Results

1.3.1 Overall Coverage, Power and Relative Bias

The reported values of coverage, power and relative bias for the sample sizes that were defined as small by the original authors are graphically displayed in boxplots in Figures 1.3, 1.4 and 1.5. On the x-axes, the different estimation methods are shown together with the number of reported values available for this estimation method. Note that coverage, power and relative bias are frequentist properties, but are still often used to evaluate and compare both frequentist and Bayesian estimation methods (Berger & Bayarri, 2004; Van Erp, Mulder, & Oberski, 2018). With the exception of 2 studies, all included simulation studies used one or multiple of these evaluation criteria and the results are combined to show the distribution of the coverage, power and relative bias levels for the varying estimation methods.[6] We divided parameters of interest into two categories: structural parameters (e.g., latent means, regression coefficients) and variance parameters (e.g., latent variances, covariances, residual variances). Note that the coverage and power for the variance parameters are not often investigated in the included studies, as those are less often the parameter of interest in substantive studies than structural parameters (see Dedrick et al., 2009) and therefore these results are discussed in text but not presented in figures. In Supplemental Table S5, the minimum, maximum and quartile values of the coverage, power and relative bias can be found for each estimation method and parameter type.

[6] The two exceptions are the studies of Farrell & Ludwig (2008) and Serang, Zhang, Helm, Steele, & Grimm (2015); they reported absolute mean bias instead of relative mean bias.

Note that the number of reported values for REML and LS is relatively small. As we have not focused explicitly on these methods, we are not able to draw any strong conclusions based on our results for REML or LS.

Coverage. In Figure 1.3, the results for the coverage of structural parameters can be found for the small sample sizes. The dashed grey lines represent the desirable coverage interval of 92.50 and 97.50 (Bradley, 1978). For the three Bayesian estimation methods, 90.97% of the values are at or above the desirable coverage of 92.50. BayesN and BayesT perform especially well: respectively 93.33% and 97.56% of coverage values are at or above 92.50. For BayesD, 64.94% are at or above 92.50. The three frequentist methods show more under-coverage than the Bayesian methods: only 52.55% of the values are above 92.50, although there are large differences between the three methods. For ML, 52.94% are at or above the desired coverage level, for REML 87.88% and for LS only 2.78%. Baldwin & Fellingham (2013) explain that coverage can be lower for frequentist methods because the sampling distribution of the parameter is assumed to be normal, an assumption which is often violated when samples are small. Hox, Moerbeek, Kluytmans, & Van de Schoot (2014) continue that because of biased standard errors for ML estimation, as a consequence of small sample sizes, ML resulted in worse coverage rates than Bayesian estimation. Using REML can improve the standard error estimates (for more information, see McNeish, 2017). This can explain why REML performs better than the other frequentist methods in terms of coverage.

The coverage levels for variance parameters for BayesT and LS are hardly investigated (number of data points = 11 and 6, respectively), and therefore no conclusions are drawn for these estimation methods based on these results. For ML and REML, 23.91% and 44.74% respectively, of the reported coverage values are at or above 92.50. Bayesian estimation performs better: for BayesN and BayesD, 65.16% and 74.00% of the reported coverage values are at or above 92.50.

Overall, Bayesian estimation lead to better coverage rates for both parameter types than the frequentist methods.

Power. In Figure 1.4, the reported power levels for the structural parameters are shown for small sample sizes. The dashed grey line represents the desirable 0.80 power level. A large part of the reported power levels of

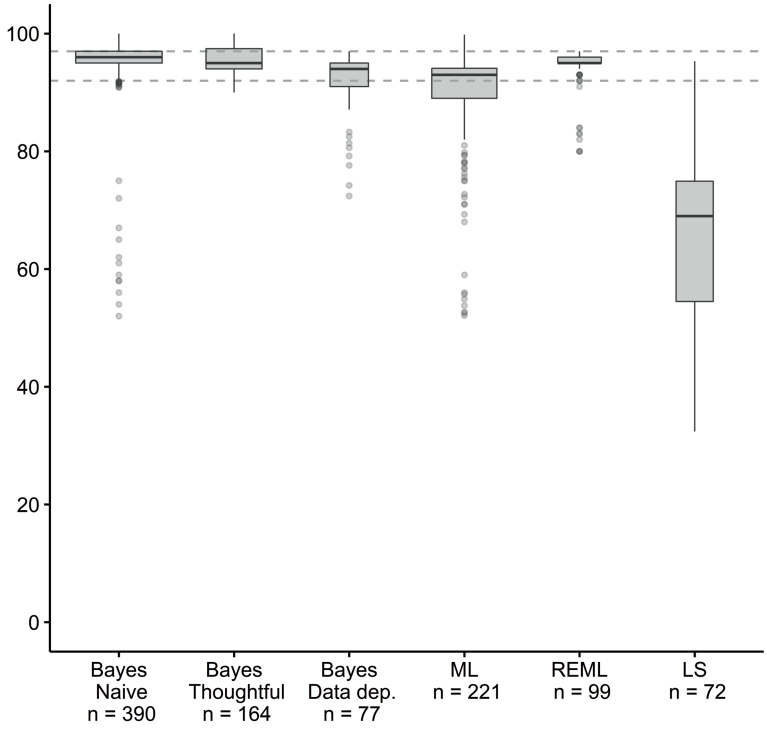


Figure 1.3: Reported coverage in the included studies for structural parameters (e.g., latent means, regression coefficients), for sample sizes defined as small by the original authors, presented for the varying estimation methods. *Note.* Dashed grey lines represent the desirable [92.50; 97.50] coverage level interval. The n represents for each estimation method the combined number of cells in the simulation designs of the included studies, that is, the amount of data points that were available. The width of the boxplots is a function of the number of data points. The boxplots are created by using the package `ggplot2` (version 2.2.1, Wickham, 2016) in R (R Core Team, 2022). The bold black line in the boxplots represent the median, the lower and upper ends of the boxplot correspond to the first and third quartiles, the whiskers are based on 1.5 times the inter-quartile range, and the circles beyond the end of the whiskers represent outliers.

the structural parameters is below 0.80. For BayesN, 85.58% are below 0.80, for BayesT 51.29%, for BayesD 78.79%, for ML 90.65%, for REML 87.20%, and for LS 87.20%. Only when BayesT was used, and thus prior information was included, power of 0.80 was reached in a substantial portion (48.71%) of the reported cases. In studies in which power levels of 0.80 or higher were reported when using BayesT, it is explained that power increased when the variance hyperparameter of the prior distribution became smaller, that is, when specific prior information is included (Miočević, MacKinnon, & Levy, 2017; Price, 2012; Van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & Loey, 2015; Zondervan-Zwijnenburg, Depaoli, Peeters, & Van de Schoot, 2019). Thus, using BayesT increased chances of reaching a power level of 0.80 or higher. For the variance parameters, the power levels are hardly investigated in the included studies (number of data points varies between 0 and 39 for the estimation methods).

Relative bias. In Figure 1.5, the relative bias for the structural parameters (Figure 1.5a) and variance parameters (Figure 1.5b) is presented for the small samples. The dashed grey lines represent the desirable ±10% level of bias (Hoogland & Boomsma, 1998). For both parameter types, the median of the distributions is within the 10% interval for all estimation methods, except the median of the distribution of LS for the structural parameters (Figure 1.5a), and the median of the distribution of ML estimation for the variance parameters (Figure 1.5b). For structural parameters, the distributions of BayesN, BayesT, BayesD, ML and REML tend to equally spread around the 10% interval, while the distribution of LS is skewed upwards. For the variance parameters, the distributions of BayesN, BayesT and LS are skewed upwards, the distribution of ML is skewed downwards, and the distributions of BayesD and REML tend to equally spread around the 10% interval. Overall, the estimation of variance parameters seems to be more problematic than the estimation of structural parameters.

For both parameter types and all estimation methods, there are outliers reported. Interestingly, the highest outliers were reported for the structural parameters, while in general the estimation of structural parameters seemed to be less problematic than the estimation of variance parameters. Note that the most extreme outliers are not visible in the boxplots, as the y-axes range

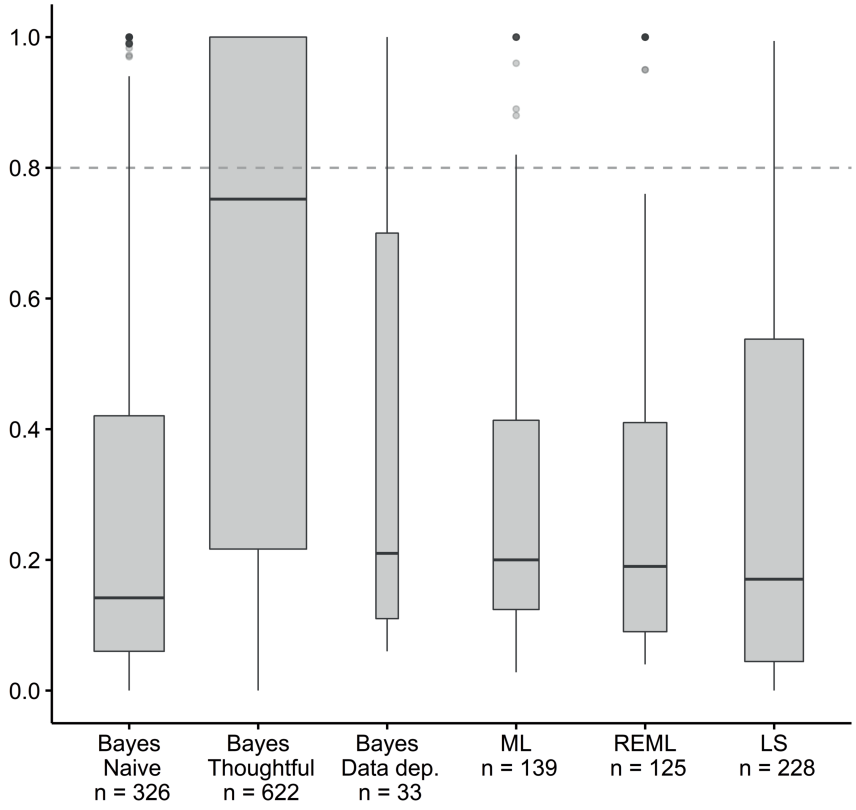


Figure 1.4: Reported power in the included studies for structural parameters (e.g., latent means, regression coefficients), for sample sizes defined as small by the original authors, presented for the varying estimation methods. *Note.* The dashed grey line represents the desirable 0.80 power level. The n represents for each estimation method the combined number of cells in the simulation designs of the included studies, that is, the amount of data points that were available. The width of the boxplots is a function of the number of data points. The boxplots are created by using the package `ggplot2` (version 2.2.1, Wickham, 2016) in R (R Core Team, 2022). The bold black line in the boxplots represent the median, the lower and upper ends of the boxplot correspond to the first and third quartiles, the whiskers are based on 1.5 times the inter-quartile, and the circles beyond the end of the whiskers represent outliers.

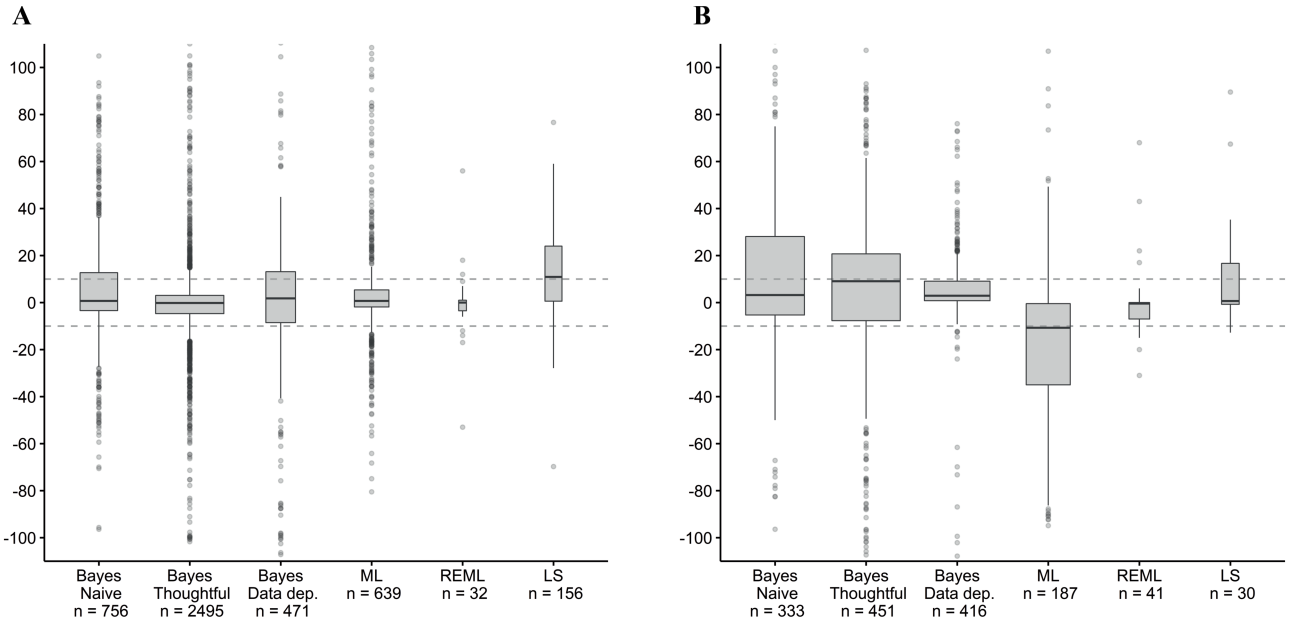


Figure 1.5: Reported relative bias in the included studies for A: structural parameters (e.g., latent means, regression coefficients), and B: variance parameters (e.g., factor variances, covariance, residual variances), for sample sizes defined as small by the original authors, presented for the varying estimation methods. *Note.* Dashed grey lines represent the desirable [−10%; +10%] relative bias interval. The n represents for each estimation method the combined number of cells in the simulation designs of the included studies, that is, the amount of data points that were available. The width of the boxplots is a function of the number of data points. The boxplots are created by using the package `ggplot2` (version 2.2.1, Wickham, 2016) in R (R Core Team, 2022). The bold black line in the boxplots represent the median, the lower and upper ends of the boxplot correspond to the first and third quartiles, the whiskers are based on 1.5 times the inter-quartile range, and the circles beyond the end of the whiskers represent outliers.

between −100% and +100% bias.[7] For BayesN, BayesT, BayesD, ML, REML and LS, respectively 54.89%, 71.74%, 41.83%, 66.82%, 78.13% and 44.23% of the reported values lie within the ±10% cutoff values for the structural parameters (Figure 1.5a). From the estimation methods, the use of REML and BayesT led to most structural parameter estimates within the ten percent boundary, followed by ML, BayesN, LS and BayesD. For the variance parameters, it is reported that for BayesN, BayesT, BayesD, ML, REML and LS, respectively 45.35%, 27.72%, 69.52%, 35.83%, 70.63% and 63.33% of the values lie within the ±10% cutoff values for the variance parameters (Figure 1.5b). From the estimation methods, the use of REML and BayesD resulted in the most variance parameter estimates within the ten percent boundary, followed by LS, BayesN, ML and BayesT. An explanation for the shift in position for BayesT is that thoughtful prior information was more often included for structural parameters than variance parameters. Note that these percentages can give a general idea of the amount of reported values within the 10% interval, but that these percentages are obviously influenced by the extreme outliers.

Overall, when looking at the median of the distributions, the performance of BayesN, BayesT, BayesD and ML is acceptable for the structural parameters. For BayesN, BayesT and ML, the performance is of poorer quality for the variance parameters, although the medians are still within the 10% interval for BayesN and BayesT. For BayesD, the performance is better for the variance parameters than the structural parameters. REML seems promising for both parameter types, although there are only 32 and 41 values reported for the structural and variance parameters respectively. Not one estimation method outperformed all others for both parameter types in terms of relative bias, when considering the percentage of reported values within the 10% cut-off values and the reported outliers.

Conclusions about overall coverage, power and relative bias. To conclude, switching to Bayesian estimation when the sample size is small, does not automatically solve small sample size problems in terms of bias. When looking at the median of the distributions, the performance of BayesN, BayesT, BayesD looks good for both parameter types, although extreme

[7]For more information about outliers, we refer to Supplemental Table S5, in which the minimum and maximum relative bias values per estimation method are reported.

outliers can occur. Higher levels of bias were found when variance parameters were estimated than when structural parameters were estimated. In terms of coverage and power, Bayesian estimation shows better results than frequentist estimation. For small samples, the desirable power level was only reached for a substantial amount of cases when BayesT was used. Bayesian estimation results in coverage mainly at or above the desired coverage level, while frequentist estimation mainly leads to values below the desired coverage level.

In the next sections, we describe the performance of Bayesian and frequentist estimation in more detail based on the results of the included simulation studies.

1.3.2 BayesN vs. Frequentist Methods

In 22 out of 32 studies, BayesN is investigated and compared to frequentist estimation. In the BayesN category, prior distributions are based on software defaults, general literature recommendations, and the use of other default priors. From the 22 studies, 5 studies reported that BayesN performs better than frequentist methods (Hox et al., 2014; Hox, Van de Schoot, & Matthijsse, 2012; Stegmueller, 2013; Tsai & Hsiao, 2008; Van Erp et al., 2018), and 3 studies reported that frequentist methods perform better (Chen, Zhang, & Choi, 2015; Depaoli & Clifton, 2015; Holtmann, Koch, Lochner, & Eid, 2016). The remaining 14 studies reported that both estimation methods performed equally or that the conclusion depended on other factors. Although one of these 14 studies reported that both frequentist and BayesN methods lead to minimal bias in the parameter estimates (Yuan & MacKinnon, 2009), 6 of 14 studies reported that both methods resulted in poor parameter estimates (Browne & Draper, 2000, 2006; Depaoli, 2013; 2 simulation studies in McNeish, 2016b; Van de Schoot et al., 2015). The remaining studies show that the conclusion depends on: the choice of the naive prior distribution (McNeish (2016a); McNeish & Stapleton (2016); e.g., McNeish & Stapleton (2016) show that BayesN with Inverse Gamma or half-Cauchy prior distributions for the variance components in a multilevel model perform better in comparison to the other BayesN option with a uniform prior distribution); the choice of the frequentist estimation method to which the BayesN is compared (Koopman,

Howe, Hollenbeck, & Sin (2015); McNeish (2016a); Miočević et al. (2017); e.g., McNeish (2016a) concludes that REML with Kenward-Roger correction performs better than ML and BayesN); or that the conclusions depend on the interest in either point estimates or interval estimates (2 simulation studies in Chen, Choi, Weiss, & Stapleton (2014)).

Despite the final conclusions of the included studies whether frequentist or BayesN estimation methods performed better, in 15 out of 22 studies that compared these estimation methods, excessively high levels of bias were reported when using BayesN. In several of these studies, there is even more bias reported with BayesN than when frequentist methods are used (see e.g., Browne & Draper, 2006; Chen et al., 2015; Depaoli & Clifton, 2015; Holtmann et al., 2016; McNeish, 2016a). As stated by McNeish (2016a) "relying on software defaults or diffuse priors with small samples can yield more biased estimates than frequentist methods." (p. 750). Besides high levels of bias, the reported levels of power were rather low (see Figure 1.4).

In 7 out of 22 studies that examined BayesN and frequentist methods, no severely biased estimates were reported when using BayesN. However, 6 of these studies focused on mediation or multilevel mediation models and did not evaluate the variance parameters (2 simulation studies in Chen et al., 2014; Hox et al., 2014; Koopman et al., 2015; Miočević et al., 2017; Yuan & MacKinnon, 2009). As shown in Figure 1.5, the variance parameters are more often problematic in terms of bias than the structural parameters. Interestingly, Tsai & Hsiao (2008) evaluated the variance parameters using Bayesian estimation with reference priors, and reported that "the Bayesian approach, particularly under the approximate Jeffreys' priors, outperforms other procedures" (p. 588). The discussion of reference priors is beyond the scope of this paper. Readers interested in reference and Jeffreys' priors are referred to Berger, Bernardo, & Sun (2009), Bernardo (1979), Jeffreys (1945), and Yang & Berger (1996).

Problematic parameters. The studies in which problematic levels of bias were reported when BayesN was used did not report problematic levels of bias for *all* parameters. Overall, the estimation of variance parameters led to substantially more problems than the estimation of structural parameters, which supports what is shown in the earlier discussed boxplot on relative bias (Figure 1.5). There were also some model specific parameters that resulted in

severely biased estimates.

In latent growth models, the highest bias was found in the estimates of the intercept variance or linear slope variance (McNeish, 2016a; Van de Schoot et al., 2015). For example, in the study by Van de Schoot et al. (2015), using BayesN (referred to as "M*plus* default priors" in Supplemental File S6), a relative bias of 84.40% is reported for the variance of the linear slope, and they report that the estimate for the intercept variance is "not even provided by M*plus* because it is too large" (p. 7).

The estimation of variance parameters in multilevel models with small samples is a well-known problem (see e.g., Gelman, 2006). This is supported by the results of the included studies. The between level variance parameters were severely biased (Browne & Draper, 2000; see e.g., Browne & Draper, 2006; Holtmann et al., 2016; Hox et al., 2012; Stegmueller, 2013) although the highest levels of relative bias were reported for the between-level covariate parameter in the study by Depaoli & Clifton (2015). The estimates for the covariate of BayesN (referred to as "noninformative (diffuse) priors" in Supplemental File S6) with a small sample size exceed the 10% cut off value in 99 of out 120 conditions (82.50%) (Depaoli & Clifton, 2015, pp. 337–344 Tables 2-7). Gelman (2006) suggested using a half-Cauchy prior distribution for the variance parameters to decrease bias. McNeish & Stapleton (2016) compared this half-Cauchy prior to an Inverse Gamma and Uniform prior for the variance components in a multilevel model (referred to as "uninformative Half-Cauchy prior", "uninformative IG prior", "uninformative U prior" in Supplemental File S6, respectively), and concluded that the half-Cauchy prior "produced the best estimates of the variance components with few clusters" (p. 12), but for the smallest number of clusters (4 clusters), the bias was "still rather high" (p. 12). For a more in-depth discussion of the half-Cauchy prior distribution, we refer to Gelman (2006) and Polson & Scott (2012).

The study by Van Erp et al. (2018) which examined a linear SEM with a mediation effect, reported problematic levels of bias for the measurement and structural intercepts. In mixture models, the recovery of class proportions was problematic when BayesN was used. The Dirichlet prior was specified for class proportions, which assumes equal class proportions in the M*plus* default settings. With a clear majority or minority class, the class proportions in

the data deviate from the ones specified by the default Dirichlet prior, and therefore resulted in very poor class proportion recovery of BayesN (Depaoli (2013); referred to as "M*plus* default noninformative priors" in Supplemental File S6).

Aside from certain parameters that require some additional attention, some other factors could also impact the performance of estimation methods, such as: categorical versus continuous data (see e.g., Holtmann et al., 2016); the strength of group differences (see e.g., Serang et al., 2015); the intra class correlations in multilevel models (see e.g., Depaoli & Clifton, 2015); the level of class separation (see e.g., Depaoli, 2012); and the number of measurement occasions (see e.g., Serang et al., 2015).

Reasons for high levels of bias. One primary culprit of the high levels of bias for the BayesN estimates is the relatively larger influence of the prior on the posterior when the sample size is small and models are complex [see e.g., Lee & Wagenmakers (2014); McNeish (2016b); Natesan (2015)]. When using naive priors, a very wide range of plausible values is specified. All values that fall within this range can be sampled during the MCMC procedure. The probability mass can therefore also lie on extreme values. This is problematic when the sample size is small, because the prior is given more relative weight than with larger samples and therefore has more impact on the posterior than it has with relatively larger sample sizes. In a complex model, there are many parameters to estimate. With a small sample size, we can expect that priors have more impact on the posterior, as the small data set is too sparse for the complexity of the model, thus making the information in the prior more impactful. The combination of the relatively large impact of the prior on the posterior and the use of default priors can result in highly biased estimates.

Furthermore, the use of improper priors could also play an important role in the cause of problematic levels of bias. Depaoli (2013) discussed that the large variance hyperparameter for the M*plus* default prior for intercepts, regression slopes and factor loadings $[N(0, 10^{10})]$ could be the reason for the highly biased parameter estimates in growth mixture models, because "the priors were acting as almost improper noninformative priors." (p. 213). Van de Schoot et al. (2015) discuss that the default hyperparameters for the Inverse Gamma distribution in M*plus* $[IG(-1, 0)]$ result in improper prior distributions, which could lead to computational problems as was pointed

out by Asparouhov & Muthén (2010). Therefore, Van de Schoot et al. (2015) recommend researchers to always use proper prior distributions instead of improper prior distributions for variance parameters, for example, use Inverse Gamma distributions with hyperparameters $(0.001, 0.001)$ which is considered to be a noninformative prior, by Van de Schoot et al. (2015, p. 9) or Inverse Gamma $(0.5, 0.5)$, which is considered to be a "very informative" prior by Van de Schoot et al. (2015, p. 9).

To conclude, using Bayesian estimation with solely naive priors does not give the desired results when sample sizes are small: it can cause extremely biased parameter estimates – even more biased than frequentist estimates – and power levels remain very low.

1.3.3 BayesT vs. Frequentist Methods

In 18 studies, BayesT was examined and compared to frequentist methods. In the BayesT condition, prior information was included for at least one of the parameters in the model, and often used in combination with flat or default priors. The investigated BayesT prior distributions in the included studies are based on (*a*) the specified population values in the simulation design; (*b*) combinations of specified population values, the literature recommendations and M*plus* default priors; (*c*) results of previous studies; and (*d*) properties of the model or knowledge of the parameter range. Especially the studies in which the priors are based on the latter two categories (*c* and *d*), can be of interest for researchers who want to apply Bayesian estimation with thoughtful priors (for the use of previous studies in prior distributions see Baldwin & Fellingham (2013) and Yuan & MacKinnon (2009); and for the use of properties of the model and knowledge of parameter range in prior distributions see Price (2012) and Yuan & MacKinnon (2009)).

From the 18 studies that compared BayesT to frequentist methods, 9 studies concluded that BayesT performed better than the frequentist methods (Depaoli & Clifton, 2015; Miočević et al., 2017; Natesan, 2015; Price, 2012; Serang et al., 2015; Van de Schoot et al., 2015; Yuan & MacKinnon, 2009; 2 simulation studies in Zondervan-Zwijnenburg et al., 2019). The other 9 studies did not report a clear preference for one of the two methods, either because BayesT and the frequentist methods performed equally well (Farrell

& Ludwig, 2008) or because the superiority of one of the two estimation methods depended on the amount or accuracy of information incorporated in the prior distributions (2 simulation studies in Depaoli, 2012, 2013; Holtmann et al., 2016; 2 simulation studies in McNeish, 2016a), the choice of the prior distributions (Baldwin & Fellingham, 2013), or the evaluation criteria and parameters of interest (Chen et al., 2014). For instance, in the two simulation studies from McNeish (2016a) it is concluded that BayesT with strong priors (referred to as "strong priors" in the latent growth model study and "strongly informative priors" in the multilevel study, in Supplemental File S6) lead to comparable results as REML with Kenward-Roger correction, and both methods perform better than BayesT with weak priors. In the two studies by Depaoli (2012), it is reported that BayesT with "tight priors" (as referred to in Supplemental File S6) performs best, followed by ML and then followed by BayesT with "weak priors" (as referred to in Supplemental File S6). Depaoli (2013) investigated 4 types of BayesT priors (referred to as "informative accurate", "weakly informative", "partial informative" and "informative and inaccurate" priors in Supplemental File S6), and concluded that only BayesT with "informative accurate priors", and BayesT with "partial knowledge priors" perform well, and that all other BayesT options and ML perform very poorly. Furthermore, Baldwin & Fellingham (2013) concluded that BayesT with Gamma priors for the variance parameters (referred to as "thoughtful priors" in Supplemental File S6) performed better than REML with Kenward-Roger correction, while REML with Kenward-Roger correction performed better than BayesT with uniform priors (referred to as "flat uniform prior" in Supplemental File S6). This shows that not only the amount of information captured in the prior distribution matters, but that also the distribution is of importance. However, in comparison to the severely biased estimates as a result of using BayesN, the bias can be extremely reduced by adjusting the parameter range without specifying a distribution that represents the prior information (Baldwin & Fellingham, 2013).

The result that BayesT performed better in general than frequentist methods is not surprising. By adding prior information, and especially when the hyperparameters of the prior distribution are centered at the population values, the posterior will give less variable and more precise results in comparison to results from frequentist methods. However, thoughtful priors

can also be specified with hyperparameters that deviate from the population values (so-called "inaccurate priors" as specified in Depaoli (2013), or "weakly/ strongly informative inaccurate priors" as specified in Holtmann et al. (2016)). Obviously, the use of these type of priors will result in worse parameter estimates compared to the result of priors with hyperparameters that are similar to the population values. However, note that the latter represent the upper-bound performance of Bayesian estimation, which is often not realistic in practice. For more details on the performance of priors that deviate from population values see for example, Depaoli (2013), Depaoli (2014), Holtmann et al. (2016), and Lee, Song & Tang (2007).

Weak vs. strong thoughtful prior distributions. In 14 studies, multiple thoughtful prior distributions are compared. These priors were obtained by varying the level of informativeness via adjusting the variance hyperparameter of the prior distribution (see e.g., Depaoli & Clifton, 2015; 2 simulation studies in Depaoli, 2012, 2013; Holtmann et al., 2016; Van de Schoot et al., 2015; 2 simulation studies in Zondervan-Zwijnenburg et al., 2019), or by adjusting both hyperparameters (2 simulation studies in McNeish, 2016a; Natesan, 2015). Other variations of thoughtful priors are obtained by varying the parameters for which a thoughtful prior was specified or by adjusting the accurateness of the prior information included in the distributions (e.g., Depaoli, 2013; Miočević et al., 2017), or finally, by varying the distribution that is specified (see e.g., Baldwin & Fellingham, 2013; Yuan & MacKinnon, 2009).

In multiple studies, it is shown that adding *weak* prior information (e.g., by specifying distributions with large variance hyperparameters), the performance can still be poor (Depaoli, 2012; Holtmann et al., 2016), probably because the admissible parameter range can still be very large. This also explains the occurrence of high levels of bias for BayesT in Figure 1.5. Even though the use of *weak* priors can still lead to biased estimates, the results are already improved in comparison to the results obtained using solely naive priors (e.g., Depaoli & Clifton, 2015; McNeish, 2016a). However, the results can further be improved by adding *stronger* prior information (e.g., Depaoli, 2012; Holtmann et al., 2016; McNeish, 2016a).

Furthermore, in mixture models, the use of BayesT in combination with a naive prior on the class proportions parameter still produces highly biased estimates (2 simulation studies in Depaoli, 2012). Depaoli (2012) concluded

that Bayesian estimation can solely be used for mixture models when "tighter priors can be placed on (...) mixture proportions and the structural model parameters" (p. 200), because it might otherwise result in higher levels of bias.

Whether a prior distribution is considered weak or strong, depends among many other factors on the parameter for which the prior is specified, and the scale of the variables in the data. To give an example of weak and strong prior distributions, we discuss the specified prior distributions in the studies of Depaoli (2012) and Holtmann et al. (2016). In both studies, normal distributions are specified $N(\mu, \sigma^2)$, where the mean hyperparameter μ equals the population value, and the variance hyperparameter σ^2 contains different values to specify the level of informativeness. First, in the study of Depaoli (2012), in which a two-factor model with two mixture classes is investigated, the variance hyperparameter for the factor loadings prior distribution was set to 100 in the "weak" condition, and set to 0.01 in the "tight" condition. A variance of 100 corresponds to a standard deviation of 10, which means that 95% of the prior distribution lies between $[-20; 20]$ when the mean hyperparameter of the distribution equals zero. A variance of 0.01 corresponds to a standard deviation of 0.1, and thus 95% of the prior distribution lies between $[-0.2; 0.2]$ when the mean hyperparameter equals zero. A second example can be found in Holtmann et al. (2016): the "weakly informative accurate priors" for the factor loadings in the two-level confirmatory factor analysis model have a variance hyperparameter of 0.2. A variance of 0.2 corresponds to a standard deviation of 0.45, and 95% of the prior distribution lies between $[-0.90; 0.90]$. The "strongly informative accurate priors" in Holtmann et al. (2016) have a variance hyperparameter of 0.01, which equals the variance used for the "tight" informative prior in Depaoli (2012).

Priors on variance parameters. In the section on 'BayesN vs. frequentist methods', it was shown that naive priors can cause high levels of bias, especially for the variance components. Seven studies that used thoughtful priors placed thoughtful priors on the variance components (Baldwin & Fellingham, 2013; Depaoli, 2012; Holtmann et al., 2016; 2 simulation studies in McNeish, 2016a; Miočević et al., 2017; Van de Schoot et al., 2015). These studies showed that using informative priors on variance parameters reduces the bias in variance estimates compared to the use of naive priors (e.g.,

Holtmann et al., 2016; McNeish, 2016a). In only four of the other studies and conditions in which naive priors were placed on the variance components in combination with thoughtful priors on other parameters in the model, the performance of variance parameters was discussed (see 2 simulation studies in Depaoli, 2012, 2013; Holtmann et al., 2016). Depaoli (2012) and Depaoli (2013) shows that naive priors on the variance parameters also in combination with informative priors on other parameters can still result in high levels of bias in mixture models (depending on the total sample size, class proportions, and level of class separation). On the other hand, Holtmann et al. (2016) conclude that the bias for variance parameters in a multilevel model was decreased when informative priors for other parameters were specified when naive priors were used for the variance parameters. This shows that when the prior distribution for one parameter is changed, it can also influence the posterior of another parameter, even when the prior distribution for a particular parameter was held constant (e.g., Holtmann et al., 2016).

Naive vs. thoughtful priors. In 8 studies, BayesN is compared to BayesT (Chen et al., 2015; Depaoli & Clifton, 2015; Holtmann et al., 2016; 2 simulation studies in McNeish, 2016a; Miočević et al., 2017; Van de Schoot et al., 2015; Yuan & MacKinnon, 2009) and all studies concluded that BayesT performed better than BayesN. There was one exception: Holtmann et al. (2016) concluded that for the two-level confirmatory factor analysis model with continuous indicators, the performance of BayesN and BayesT was comparable. For the model with categorical indicators, the performance of the "weakly/ strongly informative accurate priors" performed better than BayesN (Holtmann et al., 2016). In the other studies, BayesT was favored over BayesN regardless of other simulation conditions. For example, Yuan & MacKinnon (2009) wrote that the quality of the estimates can be improved by including prior information. Other studies in which BayesT is investigated go further in their conclusions and write that Bayes with prior information [BayesT] is *necessary* when the sample size is small. For instance, Van de Schoot et al. (2015) concluded that low levels of power and biased parameter estimates can be "solved" using Bayesian estimation with thoughtful priors (p. 1). Further, Zondervan-Zwijnenburg et al. (2019) pointed out, that to acquire reasonable power with small samples, it is necessary to use Bayesian estimation with "very specific prior information" (p. 17, and see Figure 1.3

on p. 16 in Zondervan-Zwijnenburg et al., 2019). These conclusions support the results shown in Figure 1.3, that only when Bayesian estimation is used in combination with substantial prior information, it can lead to the desired power level. When thoughtful prior distributions are placed on the parameter of interest, the power level for this particular parameter is likely to increase (Zondervan-Zwijnenburg et al., 2019), while using a naive prior on the parameter of interest - in combination with thoughtful priors for other parameters in the model - can still lead to low levels of power (McNeish, 2016a).

To conclude, when prior information centered at the population values is added to the model, it is less likely to find highly biased estimates. However, when *weak* thoughtful priors are specified, for example, because large variance hyperparameters are specified, the admissible parameter range can still be large, and therefore, the performance can still be poor (although better than when only naive priors are used). Overall, by incorporating prior information to the model, the parameter estimates improved in terms of relative bias and power.

1.3.4 BayesD vs. Frequentist Methods

In 5 studies, BayesD is compared to frequentist methods. The data dependent priors are based on ML estimates (Depaoli, 2013; McNeish, 2016b; Van Erp et al., 2018), Restricted Iterative Generalized Least Squares estimates (Browne & Draper, 2000), or BayesN estimates (Lee & Song, 2004). From these 5 studies, 3 studies reported that BayesD performed better than frequentist methods (Lee & Song, 2004; McNeish, 2016b; Van Erp et al., 2018). For example, Lee & Song (2004) favor BayesD over ML for small samples, because they found that it can even be used with samples as small as "two or three times the number of unknown parameters" (Lee & Song, 2004, p. 680). Furthermore, McNeish (2016b) reports for the investigated latent growth models, that using Bayes with data dependent priors still results in some parameter bias, but that the performance is much improved in comparison to Full Information ML or naively applying Bayes with M*plus* default priors.

The 2 remaining studies reported that both BayesD and the frequentist methods did not perform well with small samples (Browne & Draper, 2000;

Depaoli, 2013). For example, Browne & Draper (2000) summarize that Bayesian estimation [BayesD and BayesN; referred to as "gently data-determined prior", and two "diffuse Inverse Wishart priors", respectively, in Supplemental File S6] has equal or better levels of bias and coverage in comparison to two least squares frequentist estimation methods, but that "neither approach performs as well as might be hoped with small J [number of clusters]" (p. 391). The five studies that investigated BayesD yielded inconsistent results and recommendations, so it is difficult to make definitive conclusions about the performance of the BayesD approach based upon these inconclusive results.

In 3 studies, BayesD is also compared to BayesN (Browne & Draper, 2000; McNeish, 2016b; Van Erp et al., 2018). In the study of McNeish (2016b), BayesD (referred to as "data-dependent prior" in Supplemental File S6) is favored over the two BayesN priors (referred to as "noninformative proper/ improper Inverse Wishart priors" in Supplemental File S6), because it resulted in lower levels of bias and because of its ease of implementation. Browne & Draper (2000) report that both BayesD as BayesN (referred to as "gently data-determined prior" and two "diffuse Inverse Wishart priors" in Supplemental File S6) did not perform well with small samples, and Van Erp et al. (2018) concluded that, especially with small samples, all investigated methods perform very differently, and "that there is not one default prior [BayesN and BayesD; referred to as three "noninformative improper", three "vague proper", one "vague normal" and two "empirical Bayes" priors in Supplemental File S6] that performed consistently better than the other priors or than ML estimation across all parameters or outcomes." (p. 26). Depaoli (2013) compared the performance of all three Bayesian estimation methods, and concluded that Bayesian estimation with solely naive priors [BayesN; referred to as "M*plus* default noninformative priors" in Supplemental File S6] and Bayesian estimation using data dependent priors [BayesD; referred to as "data-driven informative priors" in Supplemental File S6] resulted in poor performance. Parameter estimates were well recovered only when highly informative prior distributions were used. This shows again the importance of adding prior information when Bayesian estimation is used with small samples.

1.4 Conclusion

In the current study, a systematic literature review was performed to present an overarching overview of the performance of Bayesian and frequentist estimation for structural equation models with small samples. We included 32 simulation studies in which the performance of Bayesian and frequentist estimation is compared for varying structural equation models with small sample sizes. Whereas frequentist methods can result in severely biased estimates, nonconvergence and inadmissible solutions when samples are small, Bayesian estimation can be a viable alternative. However, based on our systematic review, we strongly recommend against *naively* using Bayesian estimation to address small samples. When Bayesian estimation with solely naive priors is used, high levels of bias are reported, especially for variance parameters. This bias is often even higher than for frequentist methods, and can only be decreased by incorporating prior information, that is, using Bayesian estimation with thoughtful priors. We therefore conclude that *naively* using Bayesian estimation is not a solution for small sample problems and, what we call, *thoughtful* priors are needed. We want to encourage researchers to make well-considered decisions about *all* prior distributions when Bayesian estimation is used with small sample sizes. Therefore, in the next section, we provide recommendations on how to construct weakly thoughtful priors.

1.5 Recommendations on How to Construct Thoughtful Priors

Previous studies, meta-analysis, opinions of experts in the field, or information about the scale can be used to come up with thoughtful priors. In two included simulation studies, the authors show how they came up with thoughtful prior distributions based on previous studies (Baldwin & Fellingham, 2013; Yuan & MacKinnon, 2009). Van de Schoot et al. (2018) and Zondervan-Zwijnenburg et al. (2017) also provide useful strategies for acquiring prior information in practice. For more information on expert elicitation, we refer to O'Hagan et al. (2006).

Below, we discuss a few of many possible ways to construct thoughtful priors. We illustrate the process of selecting thoughtful priors using a mediation model (see Figure 1.6). Mediation analysis is used to evaluate the effect of an independent variable (X) on a dependent variable (Y) that is transmitted through the mediator (M). When the mediator and the outcome are continuous, the mediated effect in the single mediator model can be computed using two linear regression equations (MacKinnon, 2008):

$$M = i_2 + aX + e_2, \tag{1.1}$$

and

$$Y = i_3 + c'X + bM + e_3, \tag{1.2}$$

where i_2 and i_3 represent intercepts, a represents the effect of the independent variable on the mediator, c' represents the effect of the independent variable on the outcome controlling for the mediator, b represents the effect of the mediator on the dependent variable controlling for the independent variable, and residuals e_2 and e_3 are assumed to be normally distributed with variances $\sigma^2_{e_2}$ and $\sigma^2_{e_3}$, respectively. In Bayesian mediation analysis, the seven parameters ($i_2,\ i_3,\ a,\ c',\ b,\ e_2,\ e_3,\ \sigma^2_{e_2}$ and $\sigma^2_{e_3}$) need prior distributions. Below, we discuss hypothetical examples to construct priors for the following parameters: intercept i_2, regression coefficients a and b, and residual variance parameter $\sigma^2_{e_3}$. The examples of the prior distributions are presented in Figure 1.7, and Appendix A contains the R-code to reproduce the prior distributions.

1.5.1 Impossible and implausible parameter space

When defining priors to deal with small samples and to avoid naive priors, one could reduce the parameter space by differentiating between *impossible* parameter space – parameter values that do not receive any density mass in the prior and are prevented from occurring in the posterior, and *implausible* parameter space – values that receive very little density mass and are very improbable in the prior, but could be obtained after the prior has been updated with the data. Note that by specifying an *impossible* parameter

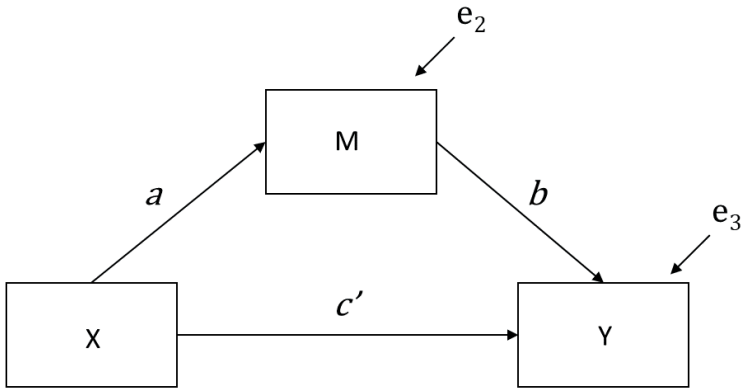


Figure 1.6: Single mediator model.

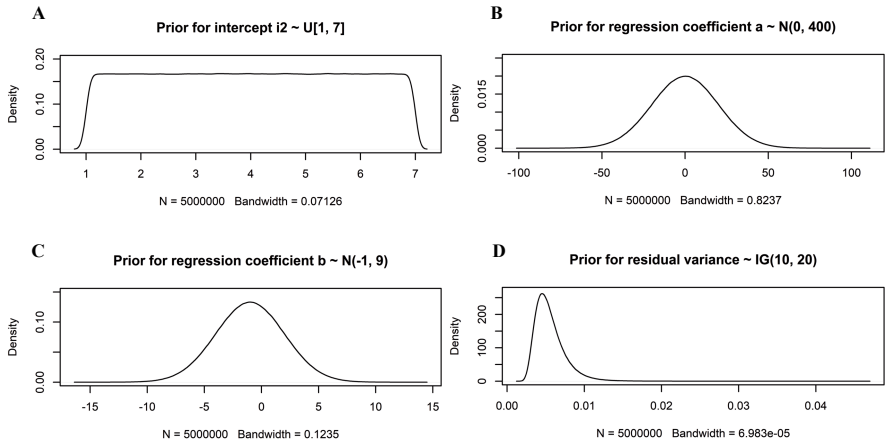


Figure 1.7: Uniform prior distribution for the intercept i_2 (see A), Normal prior distributions specified using mean and variance hyperparameters for regression coefficients a (see B) and b (see C) and Inverse Gamma prior distribution for the residual variance of Y specified using the shape (α) and scale hyperparameters (β; see D).

space (e.g. by using a Uniform or truncated-normal prior) one excludes values from the posterior – even in the case that these values do occur in the data. Therefore, we recommend to using such priors with caution and only when the excluded values are actually impossible in the data. For instance, variance parameters are often restricted to be positive, as a negative variance parameter cannot be interpreted.

When selecting a prior for the intercept of M, i_2 (see Equation (1.1)), one could specify a prior distribution based on information from the scale that is used to measure M. Suppose that a 7-point Likert scale was used to measure M. The intercept i_2 represents the value of M when X is zero (see Equation (1.1)), and given the scale of M in this case, it is impossible for M to equal any value below 1 and above 7. This is an example of an *impossible* parameter space, which can be represented by selecting a prior distribution that does not allow for values outside the range of 1 and 7, e.g., a Uniform prior distribution $U[1, 7]$ (see Figure 1.7A).

When selecting a prior for regression coefficient a (see Equation (1.1)), one could consider what constitutes an *implausible* parameter space. Suppose that in a new study where M is measured on a scale of $0 - 100$, based on the opinion of experts in the field, we expect that regression coefficients smaller than -60 and larger than 60 are highly *implausible*; that coefficients between -40 and 40 are *implausible*; and that coefficients between -20 and 20 are most plausible. Based on this information, we can compute the appropriate variance hyperparameter of the normal prior distribution. A standard deviation of 20 equals a variance hyperparameter of 400, and corresponds to a normal prior distribution in which 68% of the distribution lies between $[-20; 20]$, and 95% of the prior distribution lies between $[-40; 40]$, and 99.70% of the distribution lies between $[-60; 60]$. Based on this information, the corresponding mean hyperparameter can be computed, leading to a normal prior distribution with a mean hyperparameter of 0, and a variance hyperparameter of 400 (see Figure 1.7B). Note that although we use a normal prior distribution in the example, other types of prior distributions are also possible, depending on the software program.

1.5.2 Previous Literature

Now suppose that there is *relevant background information* about the relation between M and Y, which can be used to specify the prior for regression coefficient b (see Equation (1.2)). Let's say after performing a literature search it appears that 58% of the papers reported a negative regression coefficient, 10% reported a coefficient close to zero, and 32% of the studies reported a positive coefficient. One could create a normal prior distribution that represents these findings. For instance, a normal distribution with a mean hyperparameter of -1 and a variance hyperparameter of 9 yields these percentages (see Figure 1.7C). Note that regression coefficient b represents the effect of M on Y controlling for X. If previous literature is used to specify the prior distribution, these previous studies should have used the same scales to measure X, M, and Y as the current study, and should have been controlling for the same covariate X in the model where M predicts Y.

If the consulted literature is not an ideal source of prior information (e.g., the variables in previous studies are not the same as in the current study; or the constructs being evaluated are related, but slightly different), one can choose to make the prior less informative by increasing the variance hyperparameter. Similarly, all detected literature may suggest that the regression coefficient is negative. However, we advocate against including only negative values in the possible parameter space. Instead, in this case we recommend using a prior that allows for positive values, but makes them less probable than negative values. For examples in which expert knowledge and previous literature is used to construct priors, see Van de Schoot et al. (2018) and Zondervan-Zwijnenburg et al. (2017).

1.5.3 Variance Parameters

Selecting prior distributions for variance parameters might be less intuitive. The prior distribution that is often used for variance parameters is the Inverse Gamma distribution, which consists of two hyperparameters: α and β. To determine the values of these hyperparameters, information from a previously observed sample, a previous study or a pilot study can be used.

Hyperparameter α then equals half of the sample size of the previous study, and hyperparameter β can be computed as half of the sample size of the previous study times the variance estimate from the previous study (Gelman et al., 2014, p. 130). To illustrate, we use this method to construct the prior distribution for the residual variance of Y, $\sigma^2_{e_3}$ (see Equation (1.2)). Suppose a researcher collects pilot data from 20 participants and fits the mediation model, obtaining an estimate for the residual variance $\sigma^2_{e_3}$ of 2. The α hyperparameter will then be $0.5 \times 20 \ = 10$, and the β hyperparameter will be $(0.5 \times 20) \times 2 \ = 20$, which will yield an Inverse Gamma ($\alpha = 10, \ \beta = 20$). This Inverse Gamma distribution can now be used as a prior distribution for residual variance $\sigma^2_{e_3}$ (see Figure 1.7D).

One can increase the uncertainty in the prior by substituting a smaller value for the sample size of the previous study in the computation of the α and β hyperparameters. In case we would like to down weigh the information from the pilot study, we would encode that the sample size was below the original sample size of 20, for example 10. This yields an Inverse Gamma ($\alpha = 5, \ \beta = 10$) prior distribution with smaller hyperparameters, and therefore a less informative Inverse Gamma distribution.

1.6 Discussion and Concluding Remarks

Various sample size recommendations exist, such as: ratios in which the number of participants and number of unknown parameters (i.e., model complexity) is taken into account (e.g., Lee & Song, 2004), rules of thumb that sample sizes below 100 are in general considered too small (Kline, 2015), or that studies with sample sizes below 200 participants should be rejected from publication (Barrett, 2007; Kline, 2015) – not to mention the numerous simulation studies in which the minimum required sample size is discussed based on the simulation results for a specific model of interest (see e.g., Hox et al., 2012; Lee & Song, 2004). As shown in the current study (see Table 1.1), whether a sample size is considered to be small depends on many other factors than only the number of participants. General rules of thumb for sample sizes cannot take into account all these factors, and we should be aware that those rules of thumb are not generalizable to all situations.

A possible limitation of every systematic literature review, and thus also of the current one, is the possibility of missing a relevant study, even though we have carried out an extensive search process and have screened 3592 unique abstracts. Another limitation of our study could be that the prior distributions of all included studies are categorized into three categories, while differences exist within categories. For instance, for thoughtful priors varying levels of informativeness are studied, ranging from *weak* to *highly informative* thoughtful prior distributions centered at population values. All are allocated in the same category, while the more informative priors (centered at population values) will obviously lead to better results in terms of bias and power, than the weaker priors (centered at population values).

Based on our systematic review, we conclude that if Bayesian estimation is used to overcome small sample problems, thoughtful priors should be specified. However, the use of thoughtful priors is not a guarantee for perfectly unbiased estimates. Thoughtful prior distributions with a large variance hyperparameter, containing a large amount of uncertainty, can still yield a large admissible parameter range. They can therefore still result in poor estimates, although these estimates are likely to be an improvement over the estimates produced by Bayesian estimation with solely naive priors. Furthermore, a prior representing a high amount of certainty is only desirable when the researcher is indeed very certain about the incorporated information.

Additionally, in simulation studies, the true population values are known and therefore prior distributions can be specified so they accurately represent values of population parameters. We must bear in mind that such results show the upper-bound performance of Bayesian estimation. In empirical work, population values are obviously not known, and the specified prior distributions are therefore likely to deviate from the data. The specification of deviating (or so-called 'inaccurate') priors will evidently lead to less favorable results compared to priors containing hyperparameters similar to population values (see e.g., Depaoli, 2013; Holtmann et al., 2016). This demonstrates the relevance of investigating the impact of specified prior distributions on the posterior by performing a sensitivity analysis (see for instructions Depaoli & Van de Schoot, 2017). In addition, trace plots should always be inspected to check for spikes. They can occur when the permissible

range for a parameter is large, and detecting spikes can be a sign of the sampling of extreme values (see e.g., Depaoli & Clifton, 2015; Van de Schoot et al., 2015).

To conclude, *naively* using Bayesian estimation is not a solution for small sample problems: the specification of *thoughtful* priors is needed. We hope that the results of the current study encourage researchers to make well-considered decisions about *all* prior distributions in the model when Bayesian estimation is used with small sample sizes.

Acknowledgments

We would like to thank Naomi Schalken for her assistance in collecting the reported data on coverage, power and relative bias from all studies included in the qualitative synthesis; and Gerbrich Ferdinands for her assistance in preparing the manuscript for resubmission.

Studies included in the systematic literature review

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*(2), 151-164.

Browne, W.J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*(3), 391–420.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis, 1*(3), 473-514.

Chen, J., Choi, J., Weiss, B. A., & Stapleton, L. (2014). An empirical evaluation of mediation effect analysis with manifest and latent variables using Markov Chain Monte Carlo and alternative estimation methods. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(2), 253-262.

Chen, J., Zhang, D., & Choi, J. (2015). Estimation of the latent mediated effect with ordinal data using the limited-information and Bayesian full-information approaches. *Behavior research methods, 47*(4), 1260-1273.

Depaoli, S. (2012). Measurement and structural model class separation in mixture CFA: ML/EM versus MCMC. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(2), 178-203.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: frequentist versus Bayesian estimation. *Psychological methods, 18*(2), 186-219.

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.

Farrell, S., & Ludwig, C. J. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic bulletin & review, 15*(6), 1209-1217.

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.

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*(2), 87-93.

Hox, J. J., Moerbeek, M., Kluytmans, A., & Van De Schoot, R. (2014). Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in psychology, 5*.

Koopman, J., Howe, M., Hollenbeck, J. R., & Sin, H. P. (2015). Small sample mediation testing: misplaced confidence in bootstrapped confidence intervals. *Journal of Applied Psychology, 100*(1), 194-202.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate behavioral research, 39*(4), 653-686.

McNeish, D. (2016a). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750-773.

McNeish, D. M. (2016b). 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.

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate behavioral research, 51*(4), 495-518.

Miočeviç, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian mediation analysis for small sample research. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(5), 666-683.

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology, 6*.

Price, L. R. (2012). Small sample properties of Bayesian multivariate autoregressive time series models. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(1), 51-64.

Serang, S., Zhang, Z., Helm, J., Steele, J. S., & Grimm, K. J. (2015). Evaluation of a Bayesian approach to estimating nonlinear mixed-effects mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 202-215.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American journal of political science, 57*(3), 748-761.

Tsai, M. Y., & Hsiao, C. K. (2008). Computation of reference Bayesian inference for variance components in longitudinal studies. *Computational Statistics, 23*(4), 587-604.

Van De Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Van Loey, N. E. (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.*

Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods, 23*(2), 363-388.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological methods, 14*(4), 301-322.

Zondervan-Zwijnenburg, M., Depaoli, S., Peeters, M., & Van De Schoot, R. (2019). Pushing the limits: The performance of maximum likelihood and Bayesian estimation with small and unbalanced samples in a latent growth model. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 15*(1), 31-43.

Appendix A. R-code to reproduce the prior distributions

The R-code corresponds to the prior distributions and figures discussed in "Recommendations on how to construct thoughtful priors"

```
options(width = 60)

### Figure 1.7A, intercept i2
# Uniform prior distribution based on information of the 7-point
# Likert scale that is used to measure M
min <- 1
max <- 7
set.seed(122)
x <- runif(5000000, min = min, max = max)
plot(density(x), main = paste("Prior for intercept i2 ~ U[", min,
  ", ", max, "]", sep = ""), ylim = c(0, 0.20))

### Figure 1.7B, regression coefficient a
# Normal prior distribution based on expert knowledge on
# implausible and plausible values. Most implausible positive or
# negative value = mean ± 3 standard deviations (SDs), which
# equals ± 60 here
implausible <- 60

# To obtain the value of 1 SD, divide most implausible value by
# 3 SD
sd <- implausible/3
var <- round(sd^2, 2)

# Because we specify a normal distribution, we can find the mean
# by taking the mean of the most implausible negative and
# positive value
mean <- mean(c(-60, 60))
x <- rnorm(5000000, mean = mean, sd = sd)
```

```r
plot(density(x), main = paste("Prior for regression coefficient
  a ~ N(", mean, ", ", var,")", sep = ""), ylim = c(0, 0.025))

### Figure 1.7C, regression coefficient b
# Normal prior distribution based on studies in literature
# Values within this interval represent a null effect for this
# parameter:
int_null_lower <- -0.4
int_null_upper <- 0.4
mean <- -1
sd <- 3
var <- sd^2
set.seed(122)
x <- rnorm(5000000, mean = mean, sd = sd)

# Check the probabilities that accompany the distribution based
# on the aforementioned parameters (mean, sd, int_null_lower and
# int_null_upper). Vary the sd until you get the right
# probabilities:
x.negative<- sum(x < int_null_lower)/5000000
x.null<- sum(x < int_null_upper & x > int_null_lower)/5000000
x.positive<- sum(x > int_null_upper)/5000000

# Check the probabilities
x.negative #0.58
x.null #0.10
x.positive #0.32
plot(density(x), main = paste("Prior for regression coefficient
  b ~ N(", mean, ", ", var,")", sep = ""), ylim = c(0,0.15))

### Figure 1.7D, residual variance parameter
# Inverse Gamma prior distribution for the residual variance
# parameter
library(MCMCpack)
```

```r
# Inverse Gamma (IG) with shape and scale parameter.
# Note that in MCMC pack an IG is specified with a shape and
# rate parameter; rate = 1/scale
shape <- 10
scale <- 20
rate <- 1/scale
set.seed(122)
x <- rinvgamma(5000000, shape, rate)
plot(density(x), main = paste("Prior for residual variance
  ~ IG(", shape, ", ", scale,")", sep = ""), ylim = c(0,270))

# Less informative IG prior
shape <- 5
scale <- 10
rate <- 1/scale
set.seed(122)
x <- rinvgamma(5000000, shape, rate)
plot(density(x), main = paste("Prior for residual variance
  ~ IG(", shape, ", ", scale,")", sep = ""), ylim = c(0, 270))
```

Chapter 2

Predicting a Distal Outcome Variable from a Latent Growth Model: ML vs Bayesian Estimation

This chapter is published as: Smid, S. C., Depaoli, S., & van de Schoot, R. (2020). Predicting a Distal Outcome Variable From a Latent Growth Model: ML versus Bayesian Estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(2), 169-191. https://doi.org/10.1080/10705511.2019.1604140

Author Contributions: SS and RvdS designed the study, SS and SD further developed the study and simulation design. SS carried out the simulation study, discussed the results with SD and RvdS. SS wrote and revised the manuscript with feedback and input of SD and RvdS.

Online Data Archive and Supplementary Files: https://osf.io/ycfvg/

Abstract

Latent growth models (LGMs) with a distal outcome allow researchers to assess longer-term patterns, and to detect the need to start a (preventive) treatment or intervention in an early stage. The aim of the current simulation study is to examine the performance of an LGM with a continuous distal outcome under maximum likelihood (ML) and Bayesian estimation with default and informative priors, under varying sample sizes, effect sizes and slope variance values. We conclude that caution is needed when predicting a distal outcome from an LGM when the: (1) sample size is small; and (2) amount of variation around the latent slope is small, even with a large sample size. We recommend against the use of ML and Bayesian estimation with M*plus* default priors in these situations to avoid severely biased estimates. Recommendations for substantive researchers working with LGMs with distal outcomes are provided based on the simulation results.

2.1 Introduction

Latent growth models (LGMs) are commonly used to study developmental processes over time (Duncan, Duncan, & Strycker, 2013; Little, 2013, pp. 246–285; McArdle & Nesselroade, 2003; Meredith & Tisak, 1990). LGMs can be extended with a distal (long-term) outcome variable, which refers to a wave of assessment that occurs long after the other waves of assessment in the LGM. By estimating the regression coefficients from the latent intercept and latent slope to the distal outcome variable, researchers can examine whether someone's initial status (latent intercept) or growth rate (latent slope) can predict the distal outcome variable. Examples within the field of public health include predicting: young adult depression from conduct and emotional problems at a younger age (Koukounari, Stringaris, & Maughan, 2017); health-risking sexual behavior among young adults from adolescent substance initiation (Spoth, Clair, & Trudeau, 2014); or reading and writing problems from the development of babies with a family risk of dyslexia (Wijnen, Bree, Alphen, Jong, & Leij, 2015).

Another example of an LGM with distal outcomes is from Holgersen, Boe, Klöckner, Weisæth, & Holen (2010), who studied post-traumatic stress caused by an oil rig disaster. An LGM is analyzed with four time points closely after the oil rig disaster (one to three days; four to seven days; two weeks; and three weeks), and two distal outcome variables measured five and 27 years after the disaster. By using this model, Holgersen et al. (2010) were able to investigate whether the participants' initial status or growth rate on post-traumatic stress can predict the levels of stress five and 27 years later.

Hence, LGMs with distal outcomes allow for the assessment of longer-term patterns through the inclusion of distal outcomes. Based on the analysis of LGMs with distal outcomes, a treatment or intervention can be started sooner in order to take preventive actions. Adding a distal outcome variable to an LGM can therefore truly enhance the practical implications of a study.

In a review of the literature, no methodological or simulation studies were found examining the performance of LGMs with a distal outcome. However, if researchers base their sample size and choice for estimation method on LGM simulation results, then the expected influence of the distal outcome variable is

overlooked. Therefore, we see an important need to examine the performance of an LGM with a distal outcome, so that researchers who want to analyze such a model can rely on simulation results suitable for their model of interest. To our knowledge, this is the first simulation study examining the performance of an LGM with a distal outcome.

One important component when examining the performance of LGMs is the estimation method implemented. Much of the literature implementing LGMs represents frequentist estimation (e.g., maximum likelihood).[1] A viable alternative estimation method that is more recently established in the literature is Bayesian estimation (for examples of Bayesian LGM simulation studies see: Van de Schoot et al., 2015; Zhang, Hamagami, Wang, Nesselroade, & Grimm, 2007; Zondervan-Zwijnenburg et al., 2019). Within the Bayesian framework, prior information about parameters of the model is combined with the observed data. A Markov chain Monte Carlo (MCMC) estimation algorithm is used to obtain the posterior, which is a compromise between the data and the specified prior distributions.[2] One unique benefit of Bayesian estimation is the inclusion of prior information (e.g., Kruschke et al., 2012; Lee & Song, 2014; Van de Schoot & Depaoli, 2014). Using informative priors can lead to a decrease in estimation bias and an increase in statistical power compared to the results of frequentist methods, such as maximum likelihood estimation (see e.g., Miočević et al., 2017; Van de Schoot et al., 2015). These features are especially valuable under instances of small sample sizes. As discussed in – among many others – Gelman et al. (2014) and McNeish (2016a) and echoed in the literature review of Smid, McNeish, Miočević, & Van de Schoot (2020), the successful use of Bayesian estimation with small samples requires a thoughtful specification of priors.

2.1.1 Intended Goals and Organization of the Chapter

In the current study, we examine the performance of an LGM with a continuous distal outcome. Our interests are specifically on factors that have

[1] Maximum likelihood (ML) estimation is based on asymptotic theory, which implies that large sample sizes are required to meet the assumptions of the estimation method to obtain unbiased parameter estimates. For a conceptual explanation of ML estimation we refer to Myung (2003), and we refer to Meng & Rubin (1993) for a technical in-depth discussion.

[2] For an elaborative discussion of Bayesian estimation, we refer to, among many others: Depaoli & Van de Schoot (2017), Gelman et al. (2014), Kaplan (2014), Kaplan & Depaoli (2013), Kruschke (2015), S.-Y. Lee (2007), and Van de Schoot et al. (2014).

been shown to be important in the LGM literature, which include: estimator, sample size, the amount of variation around the latent slope, and effect size of the regression coefficients (see e.g., Hertzog, Oertzen, Ghisletta, & Lindenberger, 2008; Liu, Zhang, & Grimm, 2016; McNeish, 2016a; Van de Schoot et al., 2015; Zondervan-Zwijnenburg et al., 2019). Regarding the estimator, we are particularly interested in comparing Bayesian estimation under various prior specifications (e.g., informative priors versus diffuse default priors) to frequentist maximum likelihood estimation. This investigation will highlight 'best practice' when assessing longitudinal growth in the presence of a distal outcome.

Next, we discuss relevant simulation literature on the model, and then we introduce the (Bayesian) LGM with a distal outcome. We then describe the simulation design and discuss the results. We conclude with a discussion and recommendations for researchers working with LGMs with distal outcomes.

2.2 Previous Research on Distal Outcomes

Within the finite mixture modeling framework, distal outcomes are regularly studied. See for examples of empirical studies, Eastman, Mitchell, & Putnam-Hornstein (2016), Hipwell et al. (2016), Jiang et al. (2016), and Petras & Masyn (2010). For methodological and simulation studies see e.g., Bakk & Vermunt (2016), Bray, Lanza, & Tan (2015), Huang, Brecht, Hara, & Hser (2010), Van de Schoot, Sijbrandij, Winter, Depaoli, & Vermunt (2016), and Vermunt (2010).

Distal outcomes and covariates can impact the latent class structure within mixture models, and several methods are proposed to deal with this (Asparouhov & Muthén, 2014; Bakk, Oberski, & Vermunt, 2016; Bakk & Vermunt, 2016; Lanza, Tan, & Bray, 2013; Vermunt, 2010). Relevant to the current investigation is that adding a distal outcome increases the complexity of the model (Huang et al., 2010). A more complex model has a higher chance of non-convergence during the estimation process (Huang et al., 2010), implying a larger sample size is needed for proper estimation. Additionally, Lanza et al. (2013) discuss another factor that further complicates predicting a distal outcome variable from latent class

membership: Namely, the value of the predictor – the true class membership – is not known, but estimated in the model. There are similar concerns for the model under investigation in the current study. Akin to mixture models, the values of the predictors – the true values of the latent intercept and latent slope from the LGM – are unknown and estimated by the growth model. Furthermore, model complexity of LGMs increases when a distal outcome variable is added. We therefore expect that a relatively larger sample size is needed to circumvent convergence problems.

2.3 Latent Growth Models with a Distal Outcome

There is a rich body of simulation literature examining many different aspects of performance surrounding the LGM (see e.g., Hertzog et al., 2008; Shin, Davison, & Long, 2017; Tong & Ke, 2016). One aspect that is commonly addressed is the performance of LGM under small sample sizes (see e.g., McNeish, 2016a, 2016b, 2018; Van de Schoot et al., 2015; Zondervan-Zwijnenburg et al., 2019). Different types of LGMs have been examined in this context, however, there is no previous simulation work including distal outcomes. Simulation literature on LGMs (that do *not* include distal outcomes) has shown that small sample sizes in relation to the complexity of the model (i.e., $N < 50$, in an LGM with four time-points and two covariates) can lead to convergence problems when frequentist methods are used (see e.g., McNeish, 2016a). Furthermore, analyzing LGMs with small sample sizes can lead to biased parameter estimates and low levels of statistical power when frequentist methods are used (e.g., Van de Schoot et al., 2015; Zondervan-Zwijnenburg et al., 2019). There are similar concerns for LGMs with distal outcomes. The distal outcome variable can cause higher rates of dropouts, as the time interval between the different measurement moments is longer than for LGMs without distal outcomes.

2.3.1 The Model

Consider a general LGM with a latent intercept and a latent linear slope, as originally described by McArdle (1986), McArdle & Epstein (1987), and Meredith & Tisak (1990).[3] The LGM consists of a measurement model (2.1) and structural model (2.2):

$$\boldsymbol{y_{it}} = \boldsymbol{\eta_{I}}_i + \boldsymbol{\eta_{S}}_i \lambda_t + \boldsymbol{\varepsilon}_{it}, \tag{2.1}$$

with

$$\begin{aligned} \boldsymbol{\eta_{I}}_i &= \alpha_{I0} + \boldsymbol{\xi_{I}}_i, \\ \boldsymbol{\eta_{S}}_i &= \alpha_{S0} + \boldsymbol{\xi_{S}}_i, \end{aligned} \tag{2.2}$$

where $\boldsymbol{y_{it}}$ is the observed outcome for person i at time t, $\boldsymbol{\eta_{I}}_i$ and $\boldsymbol{\eta_{S}}_i$ respectively represent the person-specific latent intercept and latent linear slope factors, λ_t denotes the time score at time t, and $\boldsymbol{\varepsilon}_{it}$ is the person- and time-specific error term. α_{I0} is the population mean of individual intercept factor values, α_{S0} is the population mean of individual slope factor values, and $\boldsymbol{\xi_{I}}_i$ and $\boldsymbol{\xi_{S}}_i$ represent the differences between the latent factors ($\boldsymbol{\eta_{I}}_i$ and $\boldsymbol{\eta_{S}}_i$) and the population means (α_{I0} and α_{S0}).

The LGM can be extended by including a distal outcome variable (see Figure 2.1). When adding a distal outcome variable, the structural model, as shown in (2.2), is extended with:

$$\boldsymbol{\eta_{D}}_i = \alpha_{D0} + \beta_1 \boldsymbol{\eta_{I}}_i + \beta_2 \boldsymbol{\eta_{S}}_i + \boldsymbol{\xi_{D}}_i, \tag{2.3}$$

where $\boldsymbol{\eta_{D}}_i$ is the person-specific latent factor for the distal outcome, α_{D0} is the intercept of the distal outcome; that is, the population mean of the individual distal outcome variable values when $\boldsymbol{\eta_{I}}_i$ and $\boldsymbol{\eta_{S}}_i$ are zero. β_1 and β_2 are the regression coefficients representing the relations between the LGM and

[3] For an introduction into LGMs, we refer to, among many others: Curran, Obeidat, & Losardo (2010), Duncan et al. (2013), McArdle (2012), Little (2013), and Stoel, Wittenboer, & Hox (2004).

the distal outcome variable, and $\boldsymbol{\xi_{Di}}$ represents the person-specific difference between $\boldsymbol{\eta_{Di}}$ and α_{D0}. A more detailed description of all parameters in this model can be found in Appendix B.

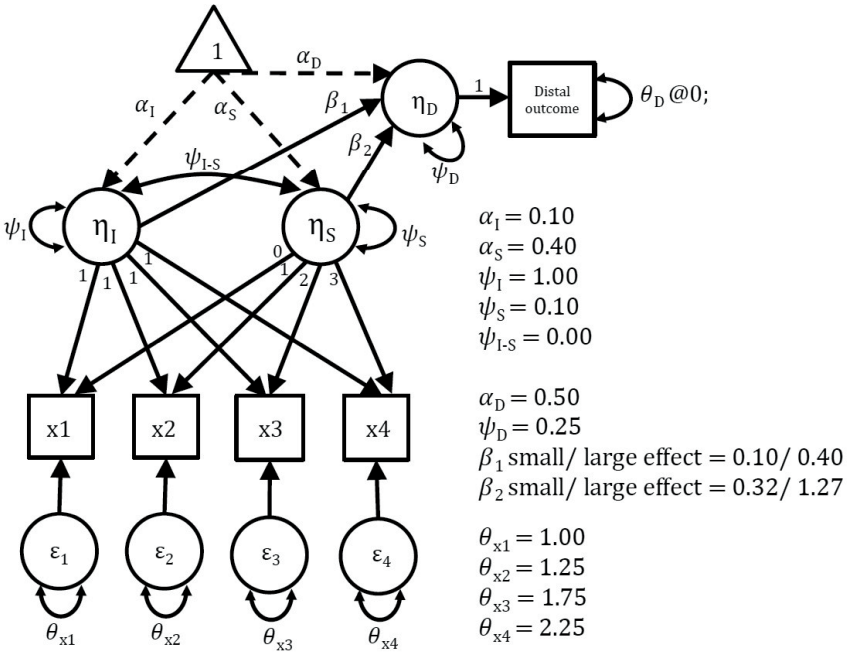


Figure 2.1: The model and population values used in the current simulation study. Note that in this figure, population values are given for a small latent slope variance ψ_s (0.10). For a large latent slope variance (1.00), the regression coefficients for β_2 are adjusted to 0.10/0.40 to still represent a small/large effect.

2.3.2 Bayesian Specification of the Model

Within the Bayesian framework, prior distributions are specified for all unknown parameters in the model. Hence, for the Bayesian LGM with a distal outcome this contains the following parameters: latent factor means $\boldsymbol{\alpha}$, regression coefficients $\boldsymbol{\beta}$, covariance matrix $\boldsymbol{\Psi}$ containing the latent factor variances and covariance $\boldsymbol{\xi}$, and matrix $\boldsymbol{\Theta}$ containing the residual variances $\boldsymbol{\varepsilon}$. We refer to these parameters as $\boldsymbol{\theta}$, which represents a vector of the unknown parameters in matrices $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}$. Hence, the prior

distributions $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\Theta})$ are denoted by $p(\boldsymbol{\theta})$.

We followed the discussion in S.-Y. Lee (2007, pp. 95–98), and adjusted the posterior distribution for the inclusion of a distal outcome here. In the posterior analysis, the observed data $\boldsymbol{Y}(y_{it}, \ldots, y_{nt})$ is augmented with the matrix of latent variables $\boldsymbol{\eta}$, resulting in the joint posterior distribution $[\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{Y}]$. The unknown parameters in $\boldsymbol{\theta}$ can be divided into two groups: $\boldsymbol{\theta_y}$, the unknown parameters in $\boldsymbol{\Theta}$ associated with the measurement model; and $\boldsymbol{\theta_w}$, the unknown parameters in $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\Psi}$ associated with the structural model. The prior distributions of the measurement model are assumed to be independent of the prior distributions of the structural model, and can therefore be seen as two different sets of prior distributions: $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}_y)p(\boldsymbol{\theta}_w)$. Hence, the likelihood is expressed by $p(\boldsymbol{Y}|\boldsymbol{\eta}, \boldsymbol{\theta}) = p(\boldsymbol{Y}|\boldsymbol{\eta}, \boldsymbol{\theta}_y)$ and $p(\boldsymbol{\eta}|\boldsymbol{\theta}) = p(\boldsymbol{\eta}|\boldsymbol{\theta}_w)$. Accordingly, the posterior distribution of the LGM with a distal outcome is given by $p(\boldsymbol{\theta}_y,\boldsymbol{\theta}_w|\boldsymbol{Y}, \boldsymbol{\eta}) \propto [p(\boldsymbol{Y}|\boldsymbol{\eta}, \boldsymbol{\theta}_y)p(\boldsymbol{\theta}_y)][p(\boldsymbol{\eta}|\boldsymbol{\theta}_w)p(\boldsymbol{\theta}_w)]$.

2.4 Simulation Design

The model of interest in the current simulation study, is an LGM with a latent intercept, latent linear slope, four time points, and one continuous distal outcome variable, as represented by (2.1), (2.2) and (2.3), and shown in Figure 2.1. The population values for this model are based on McNeish (2016a). Data sets were generated and analyzed using M*plus* version 8 (Muthén & Muthén, 1998-2017), and R version 3.4.4 via the package `MplusAutomation` version 0.7 (Hallquist & Wiley, 2017; R Core Team, 2022). The following data generation conditions were varied: sample size (3 levels), effect size (2 levels), and population values for the slope variance parameter (2 levels). These three conditions were fully crossed with each other, resulting in 12 different settings for data generation. For each of these 12 settings, 1,000 data sets were generated, and we analyzed these datasets using eight different estimation methods: (ML) estimation; Bayesian estimation with M*plus* default priors; Bayesian estimation with weak, medium, and strong informative priors centered at the population values; and Bayesian estimation with weak, medium, and strong priors deviating from the population values. Accordingly, the simulation design includes: 3

Table 2.1: Overview of the simulation design

3 sample sizes: 26 (very small), 52 (small), 325 (large)
2 effect sizes for β_1 and β_2: 0.20 (small), 0.80 (large)
2 population values for slope variance: 0.10 (small), 1.00 (large)
8 estimation settings:
• Maximum likelihood estimation (ML)
• Bayesian estimation with only M*plus* default priors (BayesDefault)
• Bayesian estimation with six sets of informative priors:
• weak, medium and strong priors centered at population values (Info Weak, Info Medium, Info Strong)
• weak, medium and strong priors deviating from population values (Deviating Weak, Deviating Medium, Deviating Strong)

(sample sizes) $\times$ 2 (effect sizes) $\times$ 2 (slope variance values) $\times$ 8 (estimation methods) = 96 cells. An overview of the simulation design can be found in Table 2.1, and the varying conditions are detailed below.

2.4.1 Conditions Simulation Design

Sample Size. Sample size was computed as a factor of the number of unknown parameters, given by: $n = d * a$, where a denotes the number of unknown parameters in the model of interest, $d = 2$ represents a very small sample, and $d = 4$ represents a small sample (as discussed in Lee & Song, 2004, p. 660). In the current study, the number of unknown parameters in the model, a, is 13. Therefore, $n = 26$ represents a *very small* sample size, and $n = 52$ a *small* sample size. We also included $n = 325$ to see how the various estimation methods perform under a *large* sample size.

Effect Size. Two different effect sizes are investigated for β_1 and β_2: a small effect size, represented by a standardized regression coefficient of 0.20; and a large effect size, represented by a standardized regression coefficient of 0.80. Supplementary file S1 shows the computation of the corresponding unstandardized regression coefficients which are used for data generation (all Supplementary files are available on the Open Science Framework: https://osf.io/ycfvg/).

Slope Variance. We investigated two levels of variation around the latent slope to examine whether this influences the prediction of the distal outcome variable in an LGM. In empirical studies, the ratio of the intercept and slope variance is often small to moderate (Hertzog et al., 2008; Liu et al., 2016), and the slope variance is usually less than 1/4 of the intercept variance (Ke & Wang, 2015). Small ratios are regularly studied in the context of simulation studies (see e.g., Bauer & Curran, 2003; Liu et al., 2016; McNeish, 2016a; Muthén & Muthén, 2002). However, larger ratios are also important to examine, as empirical research can produce any values along that continuum. In the current simulation study, the intercept variance was fixed at 1.00, and the two following levels of slope variation were examined: a small slope variation of 0.10 (ratio of 1/10), and a large slope variation of 1.00 (ratio of 1/1).

Estimation Methods. To investigate the impact of various estimation methods on the results, we compared ML to seven levels of Bayesian estimation. For ML, all M*plus* default settings were used regarding convergence (see Muthén & Muthén, 1998-2017). For the Bayesian analyses, the median point estimate of the posterior was saved, no thinning was used (i.e., thinning interval = 1), and two Markov chains were specified for each model parameter. The Gelman-Rubin potential scale reduction (PSR) factor (see e.g., Gelman et al., 2014; Gelman & Rubin, 1992) was used to assess convergence. The convergence criterion was set to 0.01 instead of the 0.05 M*plus* default, to request a stricter criterion and ensure convergence was obtained. The minimum and maximum number of iterations per chain were increased and set at $50,000$ and $150,000$, respectively. The first half of the iterations within each chain was discarded as the burn-in phase, and the remaining iterations defined the posterior. Aside from the PSR factor, convergence was also visually examined for two randomly selected data sets for each of the 12 data generating conditions. These randomly selected data sets were analyzed using the different Bayesian estimation conditions, and then trace plots for all estimated parameters were visually examined for fluctuations or other signs of non-convergence.

Prior specifications for Bayesian estimation. Here, we discuss the prior specifications for the seven Bayesian estimation settings. First, Bayesian estimation with only diffuse M*plus* default priors was used (i.e.,

BayesDefault); default priors can be found in Appendix C. The other six levels of Bayesian estimation contain informative prior distributions for five parameters in the model to mimic an achievable applied data situation, where the researcher would have information about some model parameters but not all of them. The model parameters with informed priors were: the mean of the latent intercept; the mean of the latent slope; the intercept of the distal outcome; and the two regression coefficients. M*plus* default priors were used for the remaining model parameters.

Two main types of prior distributions were specified: prior distributions that contained information similar to the population values (*informative prior conditions*), and distributions that contained information that was deviating from the population values (*deviating prior conditions*). By investigating these two types of priors, we were able to examine the upper-bound performance (i.e., when prior distributions are centered at the population values), as well as a scenario that is probably more realistic in practice (i.e., when the location of the prior distributions deviates from the population values). Under these two main categories of prior *location* (centered at the population value, and deviating from it), we investigated varying degrees of precision in the prior distributions.

Specifically, we investigated weak, medium, and strong levels of certainty by manipulating the variance hyperparameter of the prior. In order to set up these conditions, we used information we gained from the results of the condition implementing all M*plus* default prior settings. Upon obtaining the results from the default prior settings, we logged the posterior standard deviation (SD) for the five model parameters in question. We then used this value to help us set three different degrees of (un)certainty within the prior setting. Specifically, for all five parameters, a normal distribution was specified, $N(\mu, \sigma^2)$, where $\sigma^2 = 150\%,\ 100\%,\ and\ 50\%$ of the (posterior SD)^2 of the default prior setting results; these three settings created weak, medium, and strong prior distributions, respectively. The weakly informative prior had a variance hyperparameter of $1.50*(posterior\ SD)^2$, indicating it contained relatively more uncertainty (i.e., more variation). The medium informative prior used the posterior standard deviation from the default prior analysis: The variance hyperparameter was $1.0*(posterior\ SD)^2$. Finally, the strong informative prior was computed as $0.50*(posterior\ SD)^2$,

indicating it contained the most certainty out of the three conditions.

For the informative prior distributions, the mean hyperparameter (μ) of the prior distribution was set to the population value in order to center the bulk of the prior over the population value. For the deviating prior distribution conditions, the mean hyperparameter (μ) was computed such that it deviated from the true population value. Specifically, μ was specified in order that there was a five percent overlap between the informative and deviating prior distributions. For example, in the case of parameter β_2 with $n = 325$, small slope variance and small effect size, the population value was 0.32. The corresponding posterior SD produced by BayesDefault was 0.4086. Accordingly, the weakly informative prior was $N(0.32,\ (0.4086^2 * 1.5\ =\ )\ 0.25)$, the medium informative prior was $N(0.32,\ (0.4086^2 * 1\ =\ )\ 0.167)$, and the strong informative prior was $N(0.32,\ (0.4086^2 * 0.5\ =\ )\ 0.083)$. Then the deviating prior distributions were fixed to overlap with these distributions by five percent (see Figure 2.2). Consequently, the weakly deviating prior was $N(-1.642,\ 0.25)$, the medium deviating prior was $N(-1.282,\ 0.167)$, and the strong deviating prior was $N(-0.813,\ 0.083)$. This allowed us to assess the impact of priors that were slightly deviating from the population, potentially representing a setting more realistic to applied inquiries where the truth of the population is unknown.

Note that in the informative conditions, the mean hyperparameters were the same in the weak, medium, and strong distributions. While in the deviating prior conditions, the means differed for the weak, medium, and strong conditions to maintain the five percent overlap between the informative and deviating prior distributions. The variance hyperparameters were the same in the informative and deviating weak conditions; the informative and deviating medium conditions; and the informative and deviating strong conditions. For more information on the varying conditions in the simulation design and the specified prior distributions, we refer to Supplementary file S1.

2.4.2 Evaluation criteria

With small samples or complex models, convergence problems, warnings, and inadmissible parameter solutions can occur. Therefore, the number of

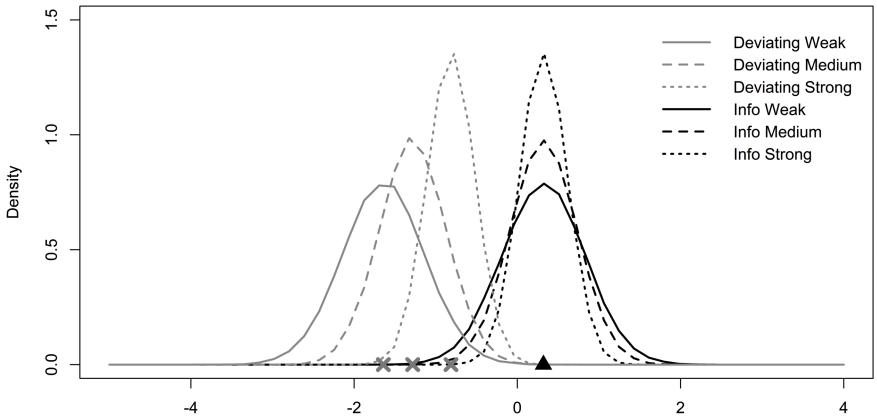


Figure 2.2: Prior distributions for regression coefficient β_2, when $n = 325$, the slope variance is small (0.10), and the effect size is small (0.20). *Note.* The three grey lines represent deviating prior distributions, which contain weak, medium and strong amounts of information respectively. The three black lines represent the weak, medium and strong prior distributions centered at population values. The triangle shows the specified population value, and the three crosses the mean hyperparameters of the three deviating prior distributions. Note that the mean hyperparameters of the three deviating priors differ, to maintain the 5% overlap with the three informative prior distributions.

completed and non-completed replications (which were highlighted by warning messages) was examined for each of the cells of the simulation design. Furthermore, for all parameters in the model, the following evaluation criteria were examined: relative mean bias, mean squared error (MSE), and coverage.

Relative mean bias was computed by $[{}^{\bar{\theta}-\theta}/_{\theta}] * 100$, where $\bar{\theta}$ denotes the average estimate across replications, and θ denotes the specified population value (Muthén & Muthén, 1998-2017). Because the population value of the covariance of the intercept and slope was zero, the relative bias could not be computed. The absolute bias (computed by $(\bar{\theta} - \theta) * 100$) is therefore reported for the covariance parameter. In interpreting parameter bias, the cutoff value of $\pm 10\%$ was used as suggested by Hoogland & Boomsma (1998). Values outside this interval represented problematic levels of bias.

The MSE was computed by $(SD)^2 + (\bar{\theta} - \theta)^2$, where SD denotes the standard deviation across replications, $\bar{\theta}$ denotes the average estimate across replications, and θ denotes the specified population value (Muthén & Muthén, 1998-2017). The MSE takes the relative bias and variability across

replications into account. Therefore, the smaller the MSE, the closer the estimated value is to the population value, across replications.

Coverage was denoted by the proportion of replications for which the 95% confidence or credibility interval contains the population value. Values between 0.925 and 0.975 are considered to represent good parameter coverage (Bradley, 1978). Values outside the interval could suggest biased standard error estimates.

Finally, for the two regression coefficients β_1 and β_2, statistical power was reported, that is, the proportion of estimates across replications that differs significantly from zero (Muthén & Muthén, 1998-2017). The preferred value for power is considered to be 0.80 (Muthén & Muthén, 2002).

2.5 Results

In this section, we focus extensively on the results of the parameters related to the distal outcome, and briefly discuss results of the other LGM parameters. The results of the medium prior distributions are very similar to either the weak or strong prior distributions, and have therefore been moved to Supplementary file S2 to conserve space.

2.5.1 Convergence and Warnings

The ML analyses did encounter convergence problems in 8.6% of the cases, when taking all 12 data generation conditions into account (out of the $12,000$ requested replications, $1,030$ of them had convergence problems). Furthermore, standard errors could not be computed in 0.075% of the cases (9 out of $12,000$ replications). The total number of completed replications under ML is shown in Table 2.2. The amount of non-completed replications when the slope variance was small is 3.11 times as high as when the slope variance was large (786 non-completed replications versus 253). Besides non-convergence errors, warnings related to the latent covariance matrix psi, and residual covariance matrix theta were given when ML was used, see Table 2.2. The total number of warnings was also higher when the slope variance was small: 2.89 times as high than when the variance of the slope

was large (total of 1,112 warnings versus 384). Warnings related to the residual covariance matrix theta were present 1.80 times more often when the slope variance was large, while warnings related to the latent variable covariance matrix psi occurred 5.44 times more often when the slope variance was small. For more information on convergence and warnings, we refer to Supplementary file S1.

From the 1,000 requested replications, the Bayesian analyses produced a 100% convergence rate, without any reported warnings. However, when visually examining trace plots to inspect if the multiple chains truly reached convergence, spikes were detected under the M*plus* default priors when small sample sizes were implemented. Spikes are extreme values sampled during MCMC, which cannot always be identified by the PSR if they are happening uniformly across the duration of the chain. For instance, for regression coefficient β_2, the trace plot showed spikes with estimates up to 4500 and down to -2000, while the population value for this parameter was 1.27.[4] With a larger sample size and large slope variance, no spikes were observed in the trace plots. Interestingly, no spikes were detected when informative and deviating priors were used in the analyses, even in the conditions with the smallest sample size. The appearance of spikes (yes/no) for the varying simulation conditions when Bayesian estimation with default priors was used, is reported in Table 2.2. Interested readers are referred to Supplementary file S1 for the trace plots with spikes for one of the examined data sets, and more details on the visual convergence checks we performed.

[4]The values of the extreme spikes for parameter β_2 correspond to the following data generation conditions: small slope variance, large effect size, $n = 26$. The replication number of the data set is 30. The spikes appeared when Bayesian estimation with default priors was used. For more information, see Supplementary file S1.

Table 2.2: Overview of the simulation process: the number of completed replications and warnings when using maximum likelihood (ML) estimation, and the occurrence of spikes when using Bayesian estimation with M*plus* default priors, for the varying simulation conditions

		Small slope variance			*Large slope variance*		
Sample Size	*Effect Size*	*Completed replications for ML*	*Warnings for ML*	*Spikes for Bayes Default*	*Completed replications for ML*	*Warnings for ML*	*Spikes for Bayes Default*
26	Small	830[a]	Total = 345, θ = 45, Ψ = 300	Yes	914	Total = 150, θ = 92, Ψ = 58	Yes
	Large	758	Total = 332, θ = 38, Ψ = 294	Yes	908	Total = 162, θ = 79, Ψ = 83	Yes
52	Small	885[c]	Total = 217, θ = 3, Ψ = 214	Yes	963[a]	Total = 31, θ = 17, Ψ = 14	Yes
	Large	787[b]	Total = 185, θ = 3, Ψ = 182	Yes	962	Total = 41, θ = 12, Ψ = 29	Yes
325	Small	986[b]	Total = 11, θ = 0, Ψ = 11	Yes	1,000	No warnings	No
	Large	968	Total = 22, θ = 22, Ψ = 0	No	1,000	No warnings	No

Note. For each of the cells, 1,000 replications were requested. The [a], [b], and [c] denote the occurrence of one, two and three non-completed replication(s), respectively, because the standard errors could not be computed. All other non-completed replications were caused by non-convergence. 'Total' shows the total number of warnings from the completed replications, reported in the M*plus* output when ML estimation was used, θ denotes the number of warnings related to the residual covariance matrix theta, and Ψ the number of problems with the latent variable covariance matrix psi. The detection of spikes (yes/ no) for BayesDefault is based on the visual assessment of trace plots for all parameters. For more details on (non-)convergence and spikes, we refer to Supplementary file S1.

2.5.2 Relative Bias

We have organized this section into subsections to promote clarity and highlight the most important patterns that emerged. First, the results of the LGM are discussed, followed by an extensive discussion of the results of the distal outcome. The results of the LGM parameters are very similar to findings from previous simulation studies and are therefore only briefly discussed in the main text.

Relative bias in the LGM. Results of the LGM parameters, can be found in Supplementary file S3. The most problematic levels of bias are found for the variance parameters of the intercept and slope, which is in line with previous LGM simulation results (see e.g., McNeish, 2016a, 2016b; Van de Schoot et al., 2015). The highest levels of bias for both parameters are reported when the slope variance was small, although the impact of the slope variance was more extreme for the variance parameter of the latent slope. Unexpected was the deterioration of both variance parameter estimates when informative priors were specified for other parameters in the model in combination with M*plus* default priors for the variance parameters (i.e., weak and strong informative prior conditions), and the improvement of the variance of the intercept parameter when deviating priors were specified for other parameters in the model (i.e., weak and strong deviating prior conditions) in comparison to the Bayesian default priors condition and ML results. For the variance parameters of the intercept and slope, ML estimation resulted in the median closest to the population value when samples were small, followed by Bayesian estimation with default priors, Bayesian estimation with informative and deviating priors (for more information, see Figures 3-4 in Supplementary file S4). This might indicate that the specified prior distributions for the variance parameters were not suitable for the current study, in combination with informative priors for other parameters in the model. This issue will be further explored and discussed in the 'Additional Exploration Priors on Variance Parameters' section, and covered in more detail in the Discussion section.

Relative bias in the distal outcome. We start this subsection with a discussion of the two regression coefficients, as these will often be the main parameters of interest in substantive studies. The section will be continued

by the description of the results of the intercept and variance of the distal outcome.

Relative bias for regression coefficients β_1 and β_2. With a small slope variance, problematic levels of bias were found when ML and the Bayesian condition with default prior settings (BayesDefault) were used, even in combination with a large sample size. Furthermore, higher levels of bias were also produced with smaller sample sizes. The specification of informative priors led to considerable improvements of the regression coefficient estimates, as can be seen in Figures 2.3-2.4. Deviating prior distributions resulted in extremely high levels of bias for the two regression coefficients. Higher levels of bias with deviating priors were associated with a small effect size. Furthermore, a counterintuitive pattern was visible for both coefficients when the slope variance was small and BayesDefault and ML were used. As can be seen in Figures 2.4A and 2.4B, the estimate of β_2 with BayesDefault was negatively biased when $n = 26$ and positively biased when $n = 325$, and vice versa for β_1 (see Figure 2.3A for β_1). The pattern disappeared when informative priors were specified (Figures 2.4A and 2.4B), and did return when deviating priors were used, when the slope variance and effect size were small (Figure 2.4E). When the slope variance was large, we did not encounter this pattern and results looked more sensible: The amount of bias decreased when the sample size increased. Note that the results for the smaller sample sizes should be interpreted with caution. When inspecting the distribution of estimates across replications, outliers were detected under small sample sizes, as well as when BayesDefault or ML were used. It is reasonable to assume that these outliers influenced the relative mean bias estimates and could have caused the counterintuitive patterns shown in Figures 2.4A and 2.4B. Therefore, boxplots are presented in Figures 2.5-2.6 to show the entire distribution of estimates across replications. Hence, the figures showing the relative mean bias should be interpreted in combination with the boxplots in which the outliers are clearly visible.

In Figures 2.5-2.6 it can be seen that when the sample size increases, the amount of outliers decreases and the distributions of estimates are closer to the true population values. Furthermore, the distributions of estimates based on Bayesian estimation with informative priors (weak and strong) were closer to the population values than when ML and BayesDefault were used.

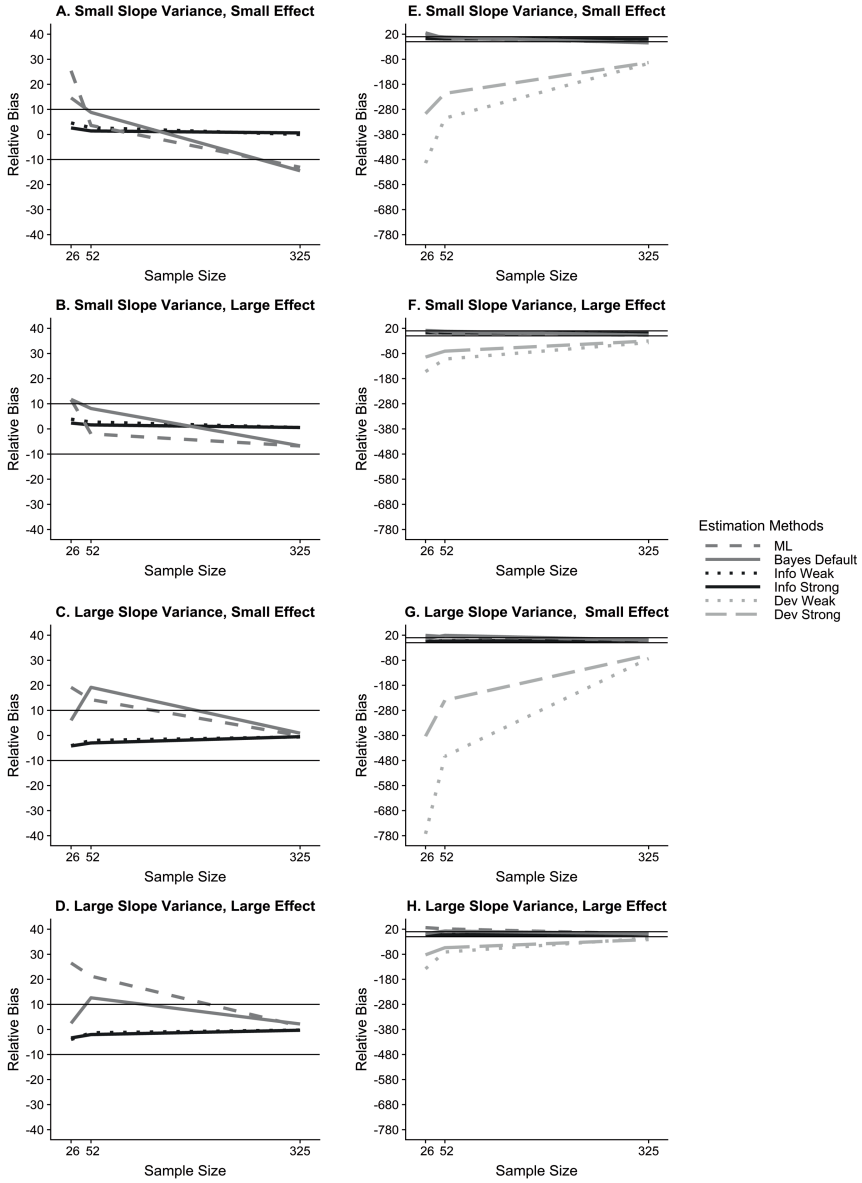


Figure 2.3: Relative bias for regression coefficient β_1, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal lines represent the ±10% interval. Subfigures at the left (A-D) present a smaller range of the y-axis to show the performance close to the ±10% boundaries, and therefore deviating prior conditions are not included here. Subfigures at the right (E-H) represent corresponding data generation conditions as A-D, but also include the deviating conditions. Note that therefore the y-axes of the A-D graphs differ from the y-axes of the E-H graphs.

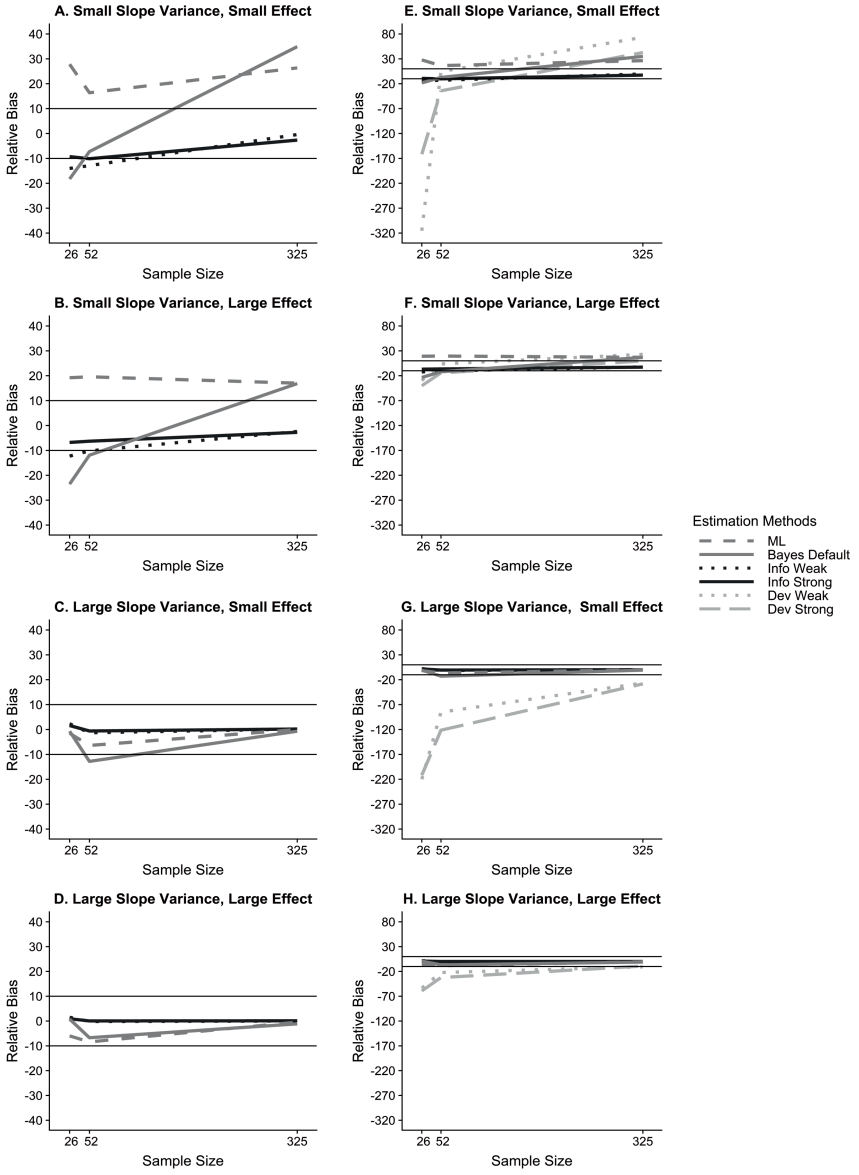


Figure 2.4: Relative bias for regression coefficient β_2, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal lines represent the ±10% interval. Subfigures at the left (A-D) present a smaller range of the y-axis to show the performance close to the ±10% boundaries, and therefore deviating prior conditions are not included here. Subfigures at the right (E-H) represent corresponding data generation conditions as A-D, but also include the deviating conditions. Note that therefore the y-axes of the A-D graphs differ from the y-axes of the E-H graphs.

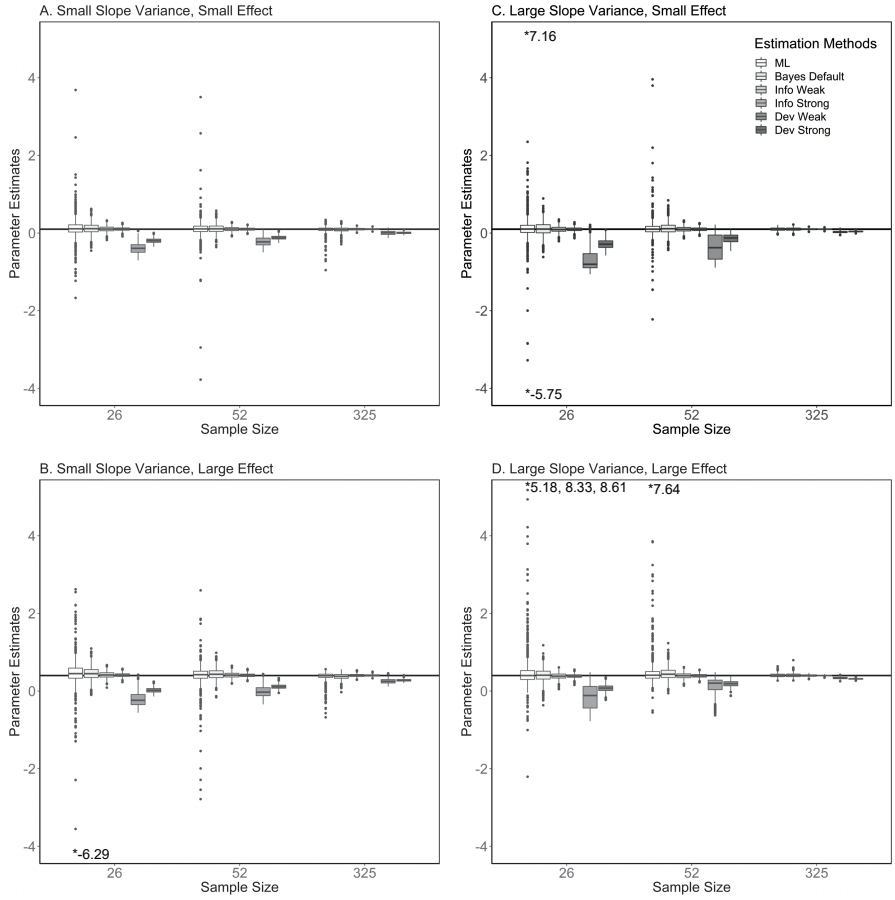


Figure 2.5: Distribution of the estimates for parameter β_1 across completed replications, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal line denotes the true population value for β_1. Outliers are displayed as black circles. Outliers outside the interval $[-4; 5]$ only occurred for ML, and are denoted by *.

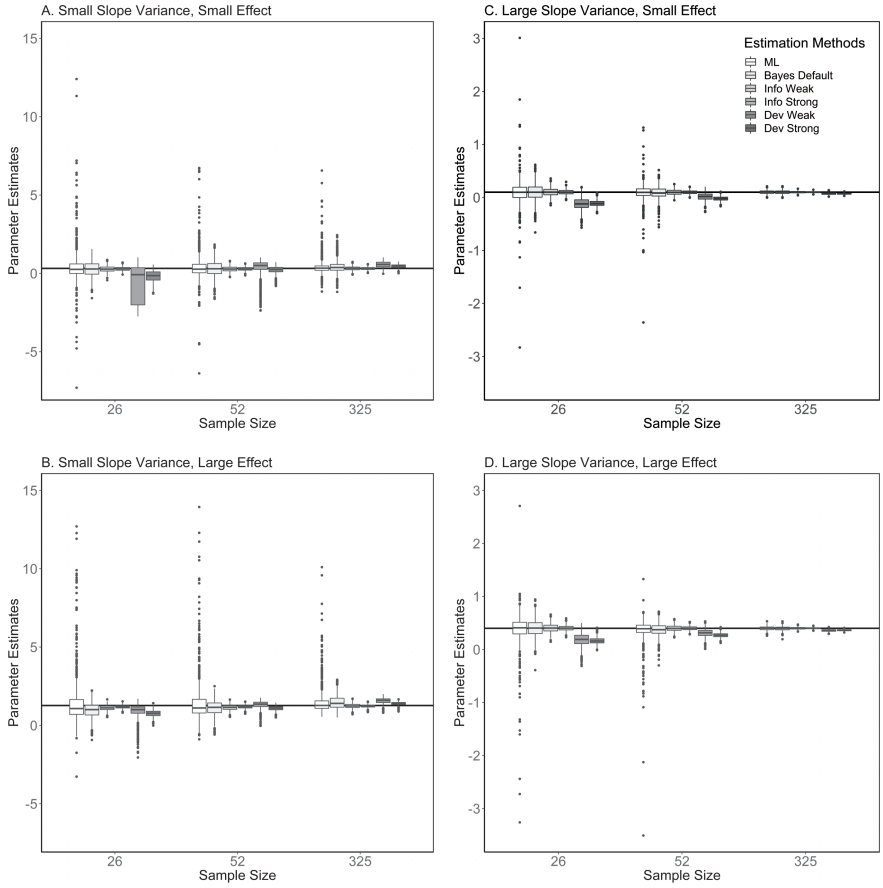


Figure 2.6: Distribution of the estimates for parameter β_2 across completed replications, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal line denotes the true population value for β_2. Outliers are displayed as black circles. Note that the range of the y-axes of subfigures A and B differs from the range of the y-axes of subfigures C and D.

The distributions of estimates based on the deviating priors (weak and strong) were clearly deviated from the population values, although less outliers did occur compared to ML and BayesDefault. Comparing β_1 and β_2, more extreme outliers were present for the estimation of β_2 (especially when the slope variance was small, see Figures 2.6A and 2.6B). For β_1, the estimation of the four data generation scenarios led to more or less similar distributions and amount of outliers.

Relative bias for intercept α_D. When ML and BayesDefault were used, too high levels of bias were only reported for the intercept of the distal outcome when the effect size was large and slope variance was small (see Figure 2.7). The estimates were biased for all sample sizes, and the counterintuitive pattern that was observed for the regression coefficients was present when the slope variance was small and the effect size was large (see Figure 2.7B). As expected, the use of informative priors improved the estimates. With a large slope variance (Figures 2.7C and 2.7D) – regardless of the effect size – the bias of the distal outcome intercept was close to zero percent when using ML, BayesDefault and the two informative prior conditions. The specification of deviating priors led to extremely biased estimates when sample sizes were small.

The boxplots in Figure 2.8 show larger ranges of distributions and more outliers when the slope variance was small (Figures 2.8A and 2.8B) compared to when the slope variance was large (Figures 2.8C and 2.8D). This indicates that the results were more stable across replications when the slope variance was large.

Relative bias for variance parameter ψ_D. For the variance of the distal outcome, biased estimates were reported for ML when the sample size was small, and this also occurred for BayesDefault when the sample size was small in combination with a small slope variance and small effect (see Figure 2.9). The estimation of the variance parameter improved when informative priors were specified for other parameters in the model and resulted in unbiased estimates. There was only one exception: When $n = 26$, the slope variance was small, and the effect was large (Figure 2.9B), the estimate was slightly biased. Deviating prior distributions, specified for other model parameters, led to increased levels of bias in comparison to ML, BayesDefault, and the informative prior conditions.

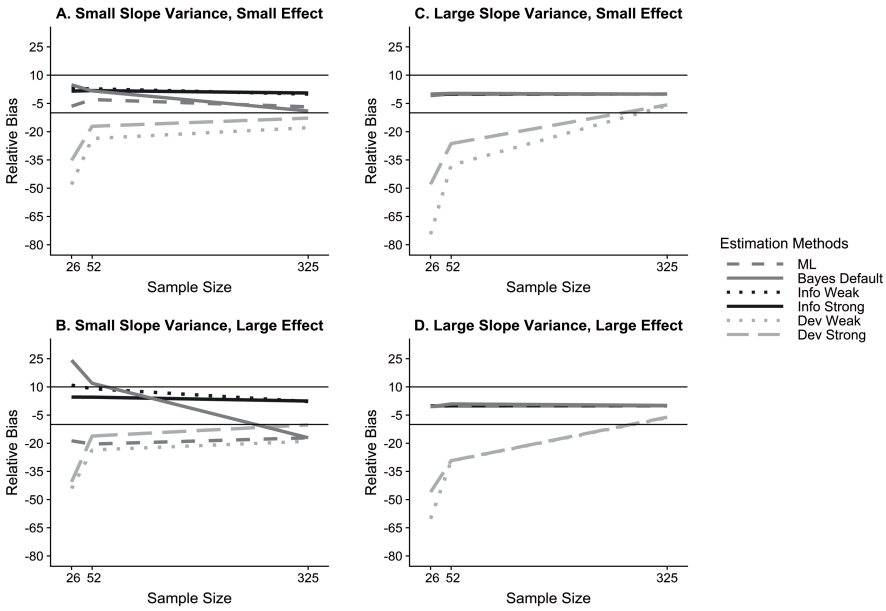


Figure 2.7: Relative bias for the intercept of the distal outcome, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal lines represent the ±10% interval.

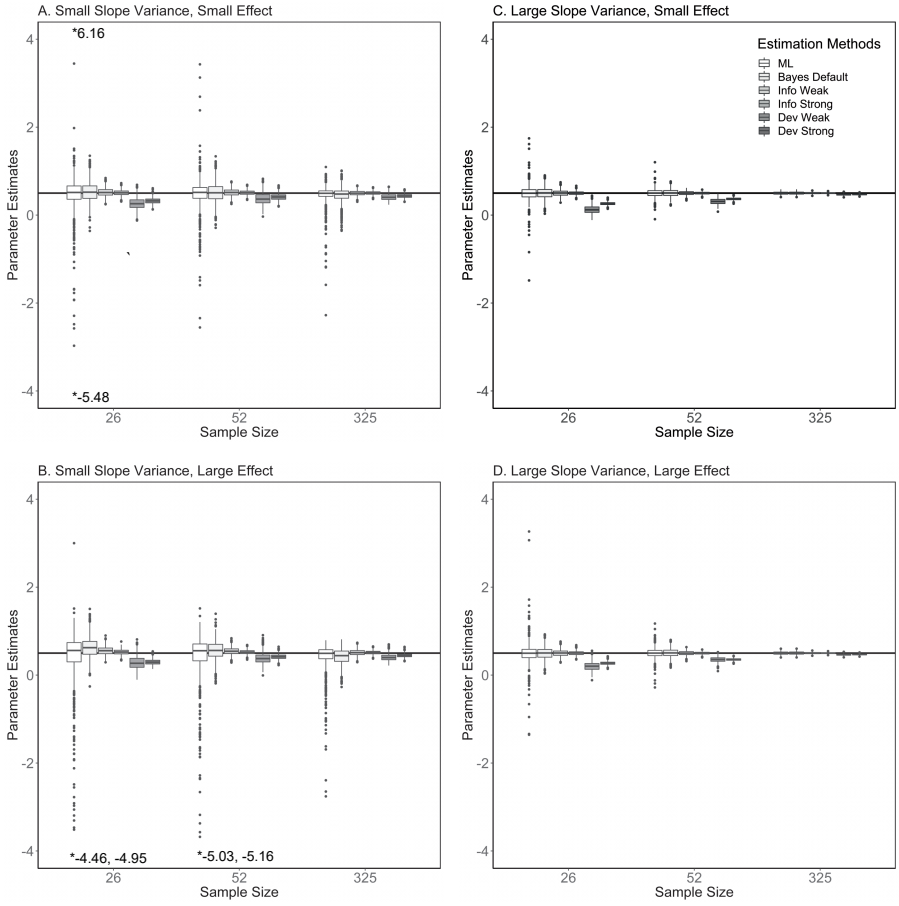


Figure 2.8: Distribution of the estimates for the intercept of the distal outcome across completed replications, under varying sample sizes, effect sizes, slope vairance values and estimation methods. *Note.* The static black horizontal line denotes the true population of 0.50 for the intercept of the distal outcome. Outliers are displayed as black circles. Outliers outside the interval [−4; 4] only occurred for ML, and are denoted by *.

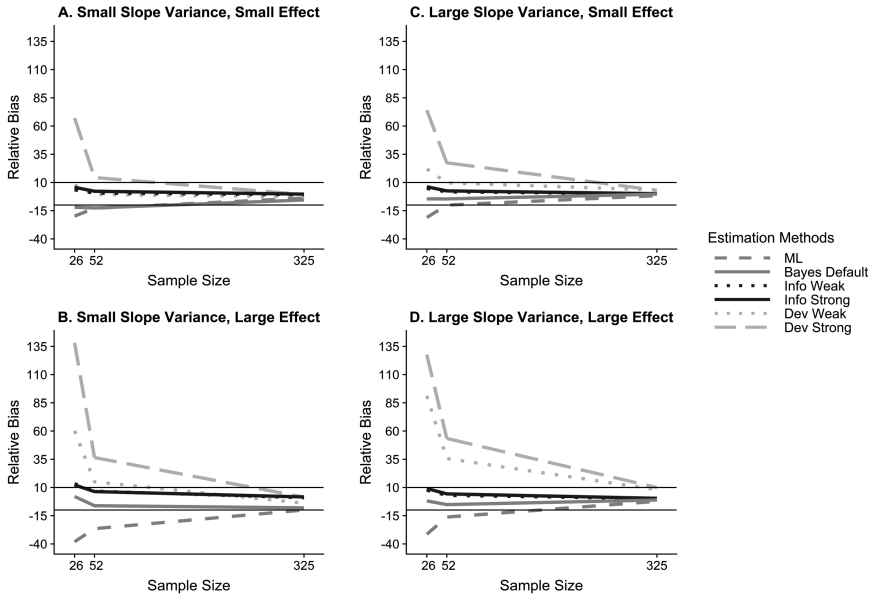


Figure 2.9: Relative bias for the variance of the distal outcome, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal lines represent the ±10% interval.

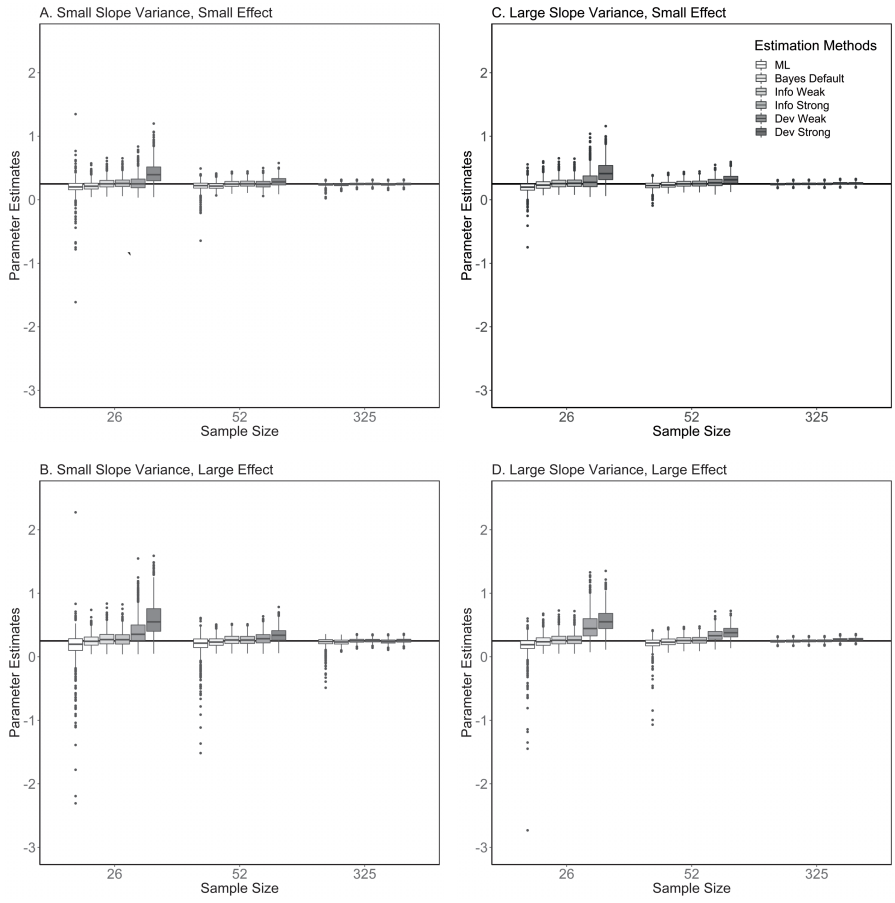


Figure 2.10: Distribution of the estimates for the variance of the distal outcome across completed replications, under varying sample sizes, effect sizes, slope variance values and estimation methods. *Note.* The static black horizontal line denotes the true population of 0.25 for the variance of the distal outcome. Outliers are displayed as black circles.

In the boxplots in Figure 2.10, it can be seen that the highest outliers were associated with ML estimation and deviating priors, when samples were small ($n = 26, 52$); and when the effect size was large.

2.5.3 Mean Squared Error

In Supplementary file S3 the mean squared error (MSE) values are shown for the varying parameters, sample sizes, effect sizes, and slope variance values. For all parameters, higher levels of MSE were associated with smaller sample sizes, the use of Bayesian estimation with weak and/or strong deviating priors, and/or ML estimation. As the MSE took into account both variability and bias of the estimates, the MSE values showed a similar pattern as the distributions of estimates shown in the boxplots in Figures 2.5, 2.6, 2.8, 2.10, and in Supplementary file S4.

2.5.4 Coverage

Results in terms of coverage can be found in Supplementary file S3. When ML was used, 21.15% of the cases showed under-coverage, but only in 5.13% of all values the coverage values were below 0.90.[5] With BayesDefault, under-coverage only occurred in 3.21%, and values below 0.90 were not obtained. Under-coverage was especially associated with smaller sample sizes. The use of informative priors never resulted in under-coverage rates, while the use of deviating priors often led to extremely low coverage values – especially for the five parameters for which deviating priors were specified. Under-coverage was found in 60.90% of the cases for weak deviating distributions, and 55.13% of the cases for the strong deviating distributions, and dramatically low coverage values were obtained down to 0.002.

ML resulted in 3.21% of the cases in over-coverage rates, while the use of BayesDefault led to over-coverage rates in 25% of the cases. The use of informative priors resulted in over-coverage rates for all situations for the five parameters for which informative priors were specified. For the other 8 parameters, over-coverage occurred in 13.54% of the cases for the weak

[5]These values represent the percentage of cases that showed under- or over-coverage from a total of 156 cases. 156 is computed as follows: 3 (sample sizes) x 2 (effect sizes) x 2 (slope variance values) x 13 (parameters) = 156.

informative priors, and in 9.38% for the strong informative priors.[6] Deviating priors hardly ever resulted in coverage rates that were too high; 2.56% of the cases for weak deviating priors yielded over-coverage, as well as 3.85% of the cases for strong deviating priors.

2.5.5 Power of the Regression Coefficients β_1 and β_2

In Table 2.3, the power rates of β_1 and β_2 are presented. With a small sample size, it was impossible to detect a small effect when ML *or* BayesDefault were used. Only in 16.7%, the power levels of ML and BayesDefault were at or above 0.80.[7] Higher levels of power were associated with a larger sample size, a large effect in the simulated data and a large slope variance. With the use of informative priors, the large effect was detected with all sample sizes for both small and large slope variances. Additionally, it became possible to detect the small effect when the largest sample size was used under informative priors. Although high levels of power were reported for the deviating priors when the effect and/ or sample size was large, the coverage was dramatically low for the corresponding estimates, especially for β_1.

2.5.6 Additional Exploration Priors on Variance Parameters

In all simulation conditions, M*plus* default priors were specified for the variance parameters (see Appendix C). After finding some unexpected results as discussed in the section: 'Relative bias in the LGM,' we explored alternate prior distributions for the variance parameters. The Inverse Wishart distribution is the default prior distribution in M*plus* for the covariance matrix of a multivariate normal distribution, which means that one prior distribution is specified for all elements in the covariance matrix (Muthén & Muthén, 1998-2017). Consequently, all elements in the covariance matrix are assigned an equal level of informativeness (e.g., Asparouhov & Muthén,

[6]Here, we used 96 cases instead of 156, because we were interested in the remaining parameters for which no informative and deviating prior distributions were specified: 2 (effects) x 2 (slope variances) x 3 (sample sizes) x 8 (parameters) = 96.

[7]Here, we used 48 cases instead of 156, because we were interested in the two regression coefficient parameters: 3 (sample sizes) x 2 (effect sizes) x 2 (slope variance values) x 2 (estimation methods) x 2 (parameters) = 48.

Table 2.3: Power of the regression coefficients β_1 and β_2

Effect Size	*Slope Variance*	*Sample Size*	*ML*	*Bayes Default*	*Info Weak*	*Info Strong*	*Dev Weak*	*Dev Strong*
Parameter: β_1								
small	small	26	**0.108**	**0.005**	**0.061**	**0.100**	**0.629**	**0.548**
		52	**0.159**	**0.016**	**0.145**	**0.274**	**0.292**	**0.253**
		325	**0.594**	**0.151**	**0.726**	0.977	**0.040**	**0.018**
	large	26	**0.089**	**0.035**	**0.053**	**0.074**	**0.642**	**0.610**
		52	**0.132**	**0.096**	**0.126**	**0.174**	**0.356**	**0.245**
		325	**0.735**	**0.710**	0.979	1.000	**0.113**	**0.324**
large	small	26	**0.534**	**0.068**	0.944	1.000	**0.085**	**0.001**
		52	**0.659**	**0.149**	0.997	1.000	**0.021**	**0.083**
		325	0.867	**0.422**	1.000	1.000	0.904	1.000
	large	26	**0.569**	**0.257**	0.937	0.998	**0.142**	**0.039**
		52	**0.766**	**0.596**	0.993	1.000	**0.332**	**0.485**
		325	0.999	1.000	1.000	1.000	1.000	1.000
Parameter: β_2								
small	small	26	**0.051**	**0.013**	**0.058**	**0.126**	**0.222**	**0.085**
		52	**0.050**	**0.035**	**0.087**	**0.174**	**0.317**	**0.168**
		325	**0.290**	**0.258**	**0.438**	**0.710**	0.811	0.916
	large	26	**0.161**	**0.020**	**0.098**	**0.154**	**0.058**	**0.106**
		52	**0.212**	**0.073**	**0.195**	**0.364**	**0.015**	**0.009**
		325	0.849	**0.794**	0.995	1.000	0.849	0.968
large	small	26	**0.294**	**0.068**	0.996	1.000	**0.592**	**0.726**
		52	**0.366**	**0.157**	0.998	1.000	0.944	0.996
		325	0.810	**0.798**	1.000	1.000	1.000	1.000
	large	26	**0.657**	**0.189**	0.986	1.000	**0.161**	**0.236**
		52	0.839	**0.359**	1.000	1.000	0.913	0.998
		325	0.999	0.994	1.000	1.000	1.000	1.000

Note. Bold values represent power rates below 0.80. ML refers to maximum likelihood estimation; BayesDefault refers to Bayesian estimation using M*plus* default priors; Info Weak and Info Strong refer to the weakly and strongly informative prior settings in the simulation design, and Dev Weak and Dev Strong to the weakly and strongly deviating prior settings.

2010). For a comprehensive discussion of the Inverse Wishart prior distribution, we refer to Schuurman, Grasman, & Hamaker (2016) and Liu et al. (2016).

As suggested by Liu et al. (2016), another option is to specify separate priors for the varying parts of the covariance matrix. This type of prior allows for separate prior distributions for each variance and covariance parameter. We followed the suggestions of Liu et al. (2016) when specifying the prior distributions for the additional exploration. An Inverse Gamma prior: $IG(0.001, 0.001)$ was specified for the variance parameters of the intercept, slope, and distal outcome. In turn, a Uniform prior, $U[-1, 1]$, was specified for the covariance of the intercept and slope. For one of the worst-case cells in the design: $n = 26$, small slope variance, small effect size, we ran $1,000$ replications for the informative and deviating prior conditions including the separate priors for the variance parameters and compared those to the existing results.

The specification of the Inverse Gamma and Uniform priors for the variance parameters resulted in an improvement of the estimates and led to sensible findings (see Supplementary file S5 for the results in terms of relative bias, and Supplementary file S6 for the boxplots of the distribution of estimates across replications). Informative priors led to a decrease in bias compared to the BayesDefault condition. In turn, deviating priors led to increased levels of bias for all four variance parameters compared to the results of informative prior conditions. For the variance of the slope and variance of the distal outcome, higher levels of bias were found for the deviating prior results compared to BayesDefault results. While for the variance of the intercept, and the covariance of the intercept and slope, the deviating prior condition showed an improvement over BayesDefault results in terms of bias. The distribution of estimates across replications (see boxplots in Supplementary file S6), showed similar patterns. The informative prior conditions resulted in distributions closer to the true population values, and less variable than the results of ML and BayesDefault. The estimates resulting from the specification of deviating priors showed more variable distributions and their medians were further away from the population value compared to the informative prior condition.

2.6 Discussion

The aim of the current study was to examine the performance of an LGM with a continuous distal outcome, under varying estimation methods, sample sizes, effect sizes, and variation around the latent slope. Caution is needed when predicting a distal outcome from an LGM latent slope when the sample size is small, or when the slope variance is small – regardless of the sample size. The use of Bayesian estimation with informative priors did improve the estimates in terms of relative bias, MSE, coverage, and power. On the other hand, the specification of priors that deviated from the population values deteriorated the results, especially when sample sizes were small.

Predicting a distal outcome variable can completely fail when there is almost no variation around the latent slope. Even a sample size of 325 is not large enough to yield unbiased regression coefficients when maximum likelihood or Bayesian estimation with default priors are used. Additionally, the prediction of the distal outcome from the latent intercept is also negatively impacted by a small variation around the latent slope, although it is less impactful when the effect size increases. Furthermore, Liu et al. (2016) associated more variation around the latent slope with higher levels of power to identify individual differences around the latent slope. A similar result was found in the current study: Higher levels of power were reported for the two regression coefficients in the large slope variance condition in comparison to the small slope variance condition.

The results of Bayesian estimation with M*plus* default priors (BayesDefault condition) in the current study, are in line with the many recent simulation studies, claiming Bayesian estimation with default priors is not preferred when sample sizes are small (see e.g., Depaoli & Clifton, 2015; Holtmann et al., 2016; McNeish, 2016a, 2016b; Shi & Tong, 2017). Spikes were detected when default priors were used in two conditions: (1) when sample sizes were small, and (2) when the slope variance and effect size were small in combination with all examined sample sizes. When prior information was incorporated, by specifying either informative or deviating prior distributions, no spikes were detected. Note that the *weakly* informative and *weakly* deviating prior conditions were still relatively informative distributions in comparison to the default diffuse priors implemented in

M*plus*. It is plausible to assume that the M*plus* default prior distributions were not informed enough to prevent spikes for the parameter estimation in the current model and are therefore likely to be the cause of high levels of bias when using Bayesian estimation with default priors.

Findings of the LGM parameters of the model were in line with the existing simulation literature on LGMs; that is, the most problematic levels of bias were detected for the variance parameters of the intercept and slope (e.g., McNeish, 2016a, 2016b; Van de Schoot et al., 2015). In McNeish (2016a), an LGM with two covariates was examined with population values and sample sizes comparable to the current study. These similarities allowed us to explore the differences in results. For instance, McNeish (2016a) showed that when using Bayesian estimation with M*plus* default priors, a sample size of 50 was sufficient to obtain an unbiased estimate for the variance parameter of the intercept. While in the current study, with an LGM with a distal outcome, a sample size of 52 still led to a biased estimate for the same parameter.

The additional exploration of the variance parameters in the model, indicated that the use of separate priors for the variance components (as suggested by Liu et al., 2016) led to sensible results, whereas the M*plus* default prior distributions for variance parameters did not. Schuurman et al. (2016) examined various specifications of the Inverse Wishart prior for the covariance matrix, and concluded that the prior settings can negatively impact the parameter estimates when the variances are close to zero. However, we decided to implement this prior in the current study based on findings of a previous studies. Based on a systematic literature review, Smid, McNeish, et al. (2020) indicated that informative priors for other parameters in the model could improve the estimates of variance parameters, when default priors were specified for the variance parameters. Depaoli (2012), Depaoli & Clifton (2015), and Holtmann et al. (2016) reported similar findings: priors on parameters in one part of the model impacted results for parameters in another part of the model. Further research is needed to examine the exact conditions under which this finding holds. Options to further explore the behavior of the variance parameters under varying prior distributions include the use of the half-Cauchy prior as suggested by Gelman (2006), or the use of reference priors as specified in Tsai & Hsiao (2008). Another option is the use of data dependent priors (Darnieder, 2011;

McNeish, 2016b), in which frequentist parameter estimates are implemented in prior distributions. One criticism of data dependent priors is that data are used twice: first to obtain frequentist parameter estimates, and second when the data is analyzed by using the data dependent priors (Darnieder, 2011). One way to avoid 'double-dipping' is the use of data-splitting techniques. For instance, in the first step, 50% of the data is analyzed using frequentist estimation, and in step two the results of step one are incorporated in prior distributions to analyze the other 50% of the data using Bayesian estimation. On the other side, as the data set needs to be split into two parts, this method is not ideal when the sample size is already small. Hence, there is no clear-cut solution as it depends on the model and the specific situation. However, based on the results of the current study, we conclude that the specification of priors for the variance parameters is of importance – regardless of the use of informative prior distributions for other parameters in the model. It is therefore necessary to further assess the Inverse Wishart (or Inverse Gamma, depending on the model) prior distribution under varying conditions in the future.

Finally, aside from the factors varied in the current simulation design, the number of time points in an LGM can also influence the performance since the number of time points directly impacts the amount of data points. Future research should therefore consider examining the potential influence of the number of time points in an LGM with a distal outcome. Another factor that should be examined is the potential impact of a categorical distal outcome instead of a continuous distal outcome. Also, the impact of adding a quadratic and/or cubic slope to the LGM could be of interest since trajectory shape is also likely to influence the impact of prior settings. Including additional slopes to the growth model increases model complexity, as there could be three to four latent factors that could be used to predict the distal outcome.

2.6.1 Recommendations for Substantive Researchers

The sample size needed to analyze an LGM with a distal outcome depends on the following four items: (1) the parameter of interest, (2) the amount of variation around the latent intercept and slope, (3) the effect size, and (4) the amount of prior information that a researcher can (or wants to) specify. For example, to predict a distal outcome from the participants' growth rate, a sample size of 26 would be sufficient to obtain an unbiased parameter estimate for the regression coefficient β_2 when using ML, BayesDefault, or Bayesian estimation with informative priors when the effect size and slope variance are large (although the statistical power was extremely low in this condition when BayesDefault was used). In contrast, a small effect and slope variance linked to a sample size of 325 would not be sufficient to obtain an unbiased estimate for β_2 when using ML or BayesDefault. In this case, only the use of informative priors could lead to an unbiased estimate when $n = 325$.

The specification of informative priors improved the estimates in terms of bias, MSE, coverage and power. Accordingly, Bayesian estimation with informative priors can be used with a smaller sample size or slope variance, and it could therefore be a solution for the analysis of data with such characteristics. However, note that informative priors represent the upper-bound performance of Bayesian estimation and not necessarily the practical application of Bayesian estimation in applied research settings. The specification of priors that deviated from the population values deteriorated the results, especially when sample sizes were small. Specifically, the deviating priors negatively influenced the estimation of the parameters for which deviating information was included, but also parameters for which no deviating information was specified.

In real life applications, it is likely to have prior distributions that at least slightly deviate from the data. One might therefore opt to choose BayesDefault instead of risking the specification of deviating priors (when comparing BayesDefault to situations when the prior deviates from the population in the current study). However, as discussed earlier, BayesDefault can lead to severely biased estimates when samples are small (see e.g., Depaoli & Clifton, 2015; Holtmann et al., 2016; McNeish, 2016a, 2016b; Shi & Tong, 2017), and is therefore hard to recommend as a viable approach.

Hence, we recommend researchers take the most careful approach possible, which entails: (1) carefully constructing prior distributions; and (2) assessing the impact and robustness of the specified priors through an extensive sensitivity analysis. For more information on how to elicit prior information (e.g., based on previous studies, meta-analyses, or knowledge of experts in the field), we refer to: O'Hagan et al. (2006); Bolsinova, Hoijtink, Vermeulen, & Béguin (2017); Zondervan-Zwijnenburg et al. (2017) and Veen, Stoel, Zondervan-Zwijnenburg, & Van de Schoot (2017); Van de Schoot et al. (2018). We also refer to Kruschke (2015, pp. 721–725) for an overview of items that should always be reported when Bayesian estimation is used, including reporting details on prior specifications. For information on how to perform a sensitivity analysis, we refer to Depaoli & Van de Schoot (2017) and Van Erp et al. (2018). An example of a sensitivity analysis in an empirical setting can be found in Van de Schoot et al. (2018).

The results of the current study further emphasize the importance of inspecting trace plots for all parameters for the appearance of spikes when using Bayesian estimation (see also Depaoli & Clifton, 2015; Van de Schoot et al., 2015). The inspection of trace plots should be a standard procedure when Bayesian estimation is used to assess whether the different chains have truly converged (see e.g., Gelman et al., 2014; Kaplan, 2014; and Lynch, 2007) – note that convergence criteria cannot always identify spikes, as we saw in the current investigation.

Although further research is needed to completely examine the performance of the Inverse Wishart prior distribution, researchers should be cautious with the use of the Inverse Wishart default prior in M*plus* for the covariance matrix. Caution is especially needed when the individual variance parameters are expected to be small (as shown by Schuurman et al., 2016). In such a situation, researchers should preferably specify separate priors as suggested by Liu et al. (2016). However, extreme caution is needed if adapting this approach, as one could easily end up with a non-positive definite matrix.

To conclude, LGMs with a distal outcome are useful to assess longer-term patterns, and to detect the need to start a (preventive) treatment or intervention in an early stage. The results of the current study showed that when predicting a distal outcome from an LGM, prudence is called for when:

(1) the sample size is small; and (2) the variance of the slope is (expected to be) small. ML and Bayesian estimation with M*plus* default prior settings should not be used in these situations to avoid severely biased estimates. A larger sample size or the specification of informative priors can help to improve the results. Note that the smaller the sample size, the larger the impact of prior distributions on the posterior, and therefore deliberate decisions about prior distributions are necessary. It is our hope that these findings help to uncover the important estimation issues tied to properly assessing the impact of distal outcomes on final model results.

Acknowledgments

We would like to thank Gerbrich Ferdinands for her assistance in preparing the manuscript for resubmission.

Appendix B. Description of the parameters in the model

Measurement model matrices:

$$\nu = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Theta = \begin{bmatrix} \theta_{x1} & & & & \\ 0 & \theta_{x2} & & & \\ 0 & 0 & \theta_{x3} & & \\ 0 & 0 & 0 & \theta_{x4} & \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

This leads to the measurement model, as given in Equation (2.4).

$$\boldsymbol{y_{it}} = \boldsymbol{\eta_{Ii}} + \boldsymbol{\eta_{Si}}\lambda_t + \boldsymbol{\varepsilon}_{it}, \tag{2.4}$$

Structural model matrices:

$$\alpha = \begin{bmatrix} \alpha_{I0} \\ \alpha_{S0} \\ \alpha_{D0} \end{bmatrix} B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_1 & \beta_2 & 0 \end{bmatrix} \Psi = \begin{bmatrix} \psi_I & & \\ \psi_S & \psi_{I-S} & \\ 0 & 0 & \psi_D \end{bmatrix}$$

This leads to the structural model, as shown in Equations (2.5) and (2.6).

$$\begin{aligned} \boldsymbol{\eta_{Ii}} &= \alpha_{I0} + \boldsymbol{\xi_{Ii}}, \\ \boldsymbol{\eta_{Si}} &= \alpha_{S0} + \boldsymbol{\xi_{Si}}, \end{aligned} \tag{2.5}$$

$$\boldsymbol{\eta_{Di}} = \alpha_{D0} + \beta_1\boldsymbol{\eta_{Ii}} + \beta_2\boldsymbol{\eta_{Si}} + \boldsymbol{\xi_{Di}}, \tag{2.6}$$

where,

$\boldsymbol{y_{it}}$ = observed outcome y for person $i(i = 1, \ldots, n)$ at time t (in simulation design: 0, 1, 2, 3)

$\boldsymbol{\eta_{Ii}}$ = random intercept factor: the expected outcome on y (here measured by $y1 - y4$) for person i at time score $\lambda_t = 0$.

$\boldsymbol{\eta_{Si}}$ = random linear slope factor: the expected outcome on y (here measured by $y1 - y4$) for person i for one unit increase in time, on the scale of λ_t.

λ_t = time score at time $t : 0, 1, 2, 3$

$\boldsymbol{\varepsilon_{it}}$ = represent individual and identically distributed measurement and time-specific errors on the y_{it} at time t, and the $\boldsymbol{\varepsilon_{it}}$ are usually assumed to be uncorrelated over time.

$\boldsymbol{\eta_{Di}}$ = random distal outcome factor: the expected outcome on d (here measured by the distal outcome variable) for person i, when taking the predictions of the latent intercept and latent slope into account.

α_{I0} = population mean of individual intercept factor values
α_{S0} = population mean of individual slope factor values
α_{D0} = population mean of individual distal outcome variable values when $\boldsymbol{\eta_{Ii}}$ and $\boldsymbol{\eta_{Si}}$ are zero, that is, the intercept of distal outcome variable

$\boldsymbol{\xi_{Ii}}$ = deviation of $\boldsymbol{\eta_{Ii}}$ from α_{I0}
$\boldsymbol{\xi_{Si}}$ = deviation of $\boldsymbol{\eta_{Si}}$ from α_{S0}
$\boldsymbol{\xi_{Di}}$ = deviation of $\boldsymbol{\eta_{Di}}$ from α_{D0}

β_1 = difference in the mean of the distal outcome factor corresponding to a one unit difference in the latent intercept factor; regression coefficient - distal outcome is regressed on latent intercept

β_2 = difference in the mean of the distal outcome factor corresponding to a one unit difference in the latent slope factor; regression coefficient - distal outcome is regressed on latent slope

Formulas and interpretation based on Masyn, Petras, & Lu (2014) and Duncan et al., (2013, pp. 56–62).

Appendix C. M*plus* default priors

- Mean latent intercept, mean latent slope, and intercept distal outcome: $N(0, 10^{10})$
- Regression coefficients: $N(0, 10^{10})$
- Variances Intercept, Slope, and Covariance Intercept-Slope: $IW(0, -3)$
- Variance Distal outcome: $IG(-1, 0)$
- Residual variances: $IG(-1, 0)$

Chapter 3

Twostep Modeling and Factor Score Regression vs Bayesian Estimation with Informative Priors

This chapter is published as Smid, S. C., & Rosseel, Y. (2020). SEM with small samples: Twostep modeling and factor score regression versus Bayesian estimation with informative priors. In R. Van de Schoot & M. Miočeviç (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*: Routledge.

Author Contributions: SS and YR designed the study; SS performed the simulation study and interpreted the results with feedback from YR; SS wrote the chapter and YR gave feedback on the written work.

Online Data Archive and Supplementary Files: https://osf.io/bam2v/

Abstract

Two promising frequentist methods to analyze SEMs with small samples are twostep modeling and factor score regression. Using a simulation study, we investigated those methods - under varying sample sizes - and compared them to maximum likelihood estimation and Bayesian estimation with default and informative priors. We conclude that with small samples, all frequentist methods showed signs of breaking down (in terms of non-convergence, negative variances, extreme parameter estimates), as well as the Bayesian condition with default priors (in terms of mode-switching behavior). When increasing the sample size is not an option, we recommend using Bayesian estimation with informative priors. However, results should be interpreted with caution, because of the large influence of the prior on the posterior with relatively small samples. When researchers prefer not to include prior information, twostep modeling or factor score regression are recommended, as those led to higher convergence rates without negative variances, more stable results across replications and less extreme parameter estimates than maximum likelihood estimation with small samples.

3.1 Introduction

Bayesian estimation is regularly suggested as a beneficial method when sample sizes are small, as pointed out by systematic literature reviews in many fields, such as: organizational science (Kruschke, 2010); psychometrics (Rupp et al., 2004); health technology (Spiegelhalter, Myles, Jones, & Abrams, 2000); epidemiology (Rietbergen et al., 2017); education (König & Van de Schoot, 2017); medicine (Ashby, 2006); and psychology (Van de Schoot, Winter, et al., 2017). Similarly, many simulation studies have shown the advantages of applying Bayesian estimation to address small sample size issues for structural equation models (SEMs), instead of using frequentist methods (see e.g., Depaoli, 2013; Muthén & Asparouhov, 2012; Stegmueller, 2013; Van de Schoot et al., 2015; Van Erp et al., 2018). However, as discussed in McNeish (2016a) and echoed in the systematic literature review of Smid, McNeish, et al. (2020), the use of Bayesian estimation with only diffuse default priors can cause extremely biased estimates when samples are small. The specification of informative priors is therefore required when Bayesian estimation is used with small samples.

Besides using Bayesian estimation with informative priors, there are also options for analyzing SEMs with small samples within the frequentist framework. Many studies have shown that the use of maximum likelihood (ML) estimation with small samples can result in convergence problems, inadmissible parameter solutions and biased estimates (see e.g., Boomsma, 1985; Nevitt & Hancock, 2004). Two newly introduced and promising frequentist methods to analyze SEMs with small samples are twostep modeling (twostep) and factor score regression (FSR). A recent development is the implementation of twostep and FSR in the accessible software `lavaan` (Rosseel, 2012), as discussed in Rosseel (2020).[1] In twostep modeling, the measurement models for the latent variables are estimated separately as a first step. As a second step, the remaining parameters are estimated while the parameters of the measurement models are kept fixed to their estimated values. Twostep modeling originates from work of Burt (1976) and Anderson

[1] Twostep modeling and FSR are both variants of the Structural-after-Measurement (SAM) approach in `lavaan`. In a nutshell, 'twostep' is gobal SAM, and 'fsr' is local SAM. For more information about SAM, we refer to (Rosseel & Loh, 2022). The original functions as used in the simulation study are still available via lavaan:::twostep() and lavaan:::fsr().

& Gerbing (1988), and more recent work can be found in the latent class literature (e.g., Bakk et al., 2014). In FSR, each latent variable in the model is replaced by factor scores and subsequently path analysis or regression analysis is ran using those factor scores. Recent developments in FSR can be found in studies of Croon (2002), Devlieger, Mayer, & Rosseel (2016), Devlieger & Rosseel (2017), Hoshino & Bentler (2013), and Takane & Hwang (2018).

No simulation studies were found in which twostep and FSR are compared to Bayesian estimation. Therefore, the goal of this chapter is to examine the performance of the following estimation methods under varying sample sizes: twostep, FSR, ML estimation, and Bayesian estimation with three variations in the specification of prior distributions. The remainder of the chapter is organized as follows: Next, the statistical model will be discussed, as well as software details, the simulation conditions, and evaluation criteria. Then, results of the simulation study will be described. We end the chapter with a summary of the results, and recommendations on when to use which estimation method in practice.

3.2 Simulation Design

3.2.1 Statistical Model

The model of interest in this simulation study is a SEM in which latent variable X is predicting latent variable Y, see Figure 3.1. Both latent variables are measured by three continuous indicators. The model and population values are similar to the model discussed in Rosseel & Devlieger (2018). The parameter of interest in the current chapter is the regression coefficient β. The standardized regression coefficient, β^Z, is 0.243, which can be considered a small effect according to Cohen (1988).

3.2.2 Software Details

Data sets were generated and analyzed in R version 3.4.4. (R Core Team, 2022), using packages `lavaan` version 0.6-1 (Rosseel, 2012) for the analyses of

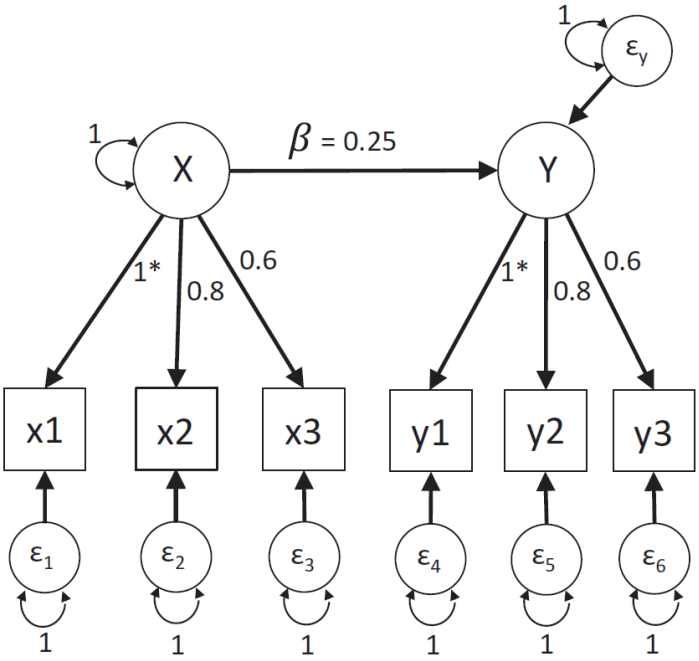


Figure 3.1: The model and unstandardized population values used in the simulation study. For scaling, the first factor loading for each factor is fixed to 1 (denoted by 1* in the figure), and the means of the latent variables are fixed to zero (not shown in the figure)

twostep, FSR and ML; and `blavaan` version 0.3-2 (Merkle & Rosseel, 2018) for the analyses of the Bayesian conditions. Example code of the analyses using the six estimation methods can be found in supplemental file S1. All simulation code and supplemental files are available online (https://osf.io/bam2v/).

Six levels of sample size were examined, and for each sample size, 1,000 data sets were generated according to the model and population values shown in Figure 3.1. Each generated data set was analyzed using six estimation methods. Accordingly, a total of 6 (sample size) x 6 (estimation methods) = 36 cells were investigated in the simulation design.

3.2.3 Simulation Conditions

Six levels of sample size are studied: $10, 20, 50, 100, 250$ *and* 500 to investigate how sample size influences the performance of the varying estimation methods. For the current model, sample sizes of 10 and 20 are extremely small. A sample size of 50 is considered small, and sample sizes of 100 and 250 are considered medium. The sample size of 500 is considered large and included as a benchmark.

Six estimation methods are considered in the current study. Three frequentist estimation methods: twostep, FSR, and ML; and Bayesian estimation with three types of prior specifications. For the three frequentist methods, all default settings of the `lavaan` package were used. For the default settings, see the help page for lavOptions() in the `lavaan` package. For the Bayesian methods, we used 4 chains instead of the 2 default chains. In terms of convergence, we used the Potential Scale Reduction (PSR) factor, set it to a stricter criterion of 1.01, and used the following minimum number of iterations: a fixed burn in period of $10,000$ iterations (specified in `blavaan` by adapt $= 2,000$, burnin $= 8,000$), and for the sampling period $20,000$ iterations (specified in `blavaan` by sample $= 20.000$).[2] As an additional check, we visually assess convergence for two randomly selected data sets for each of the sample sizes and the Bayesian conditions (2 data sets x 6 sample sizes x 3 Bayesian conditions = 36 cases), by inspecting the traceplots for all parameters.

Three variants of prior specifications were examined, and all priors were specified for unstandardized parameters: BayesDefault, BayesInfoI, and BayesInfoII, see Table 3.1. The BayesDefault condition refers to a naive use of Bayesian estimation, where only `blavaan` default priors are used. The BayesInfoI and BayesInfoII conditions refer to research situations where weakly prior information is available. In BayesInfoI, weakly informative priors are specified for the factor loadings, and `blavaan` default priors are specified for the remaining parameters. In BayesInfoII, weakly informative priors are used for both the factor loadings *and* regression coefficient β, in

[2] When the PSR criterion is not reached after the specified minimum number of iterations, the number of iterations is automatically increased until the PSR criterion is met. We adjusted the `blavaan` default for the maximum time that the software uses to increase the amount of iterations to "24 hours" instead of the default "5 minutes".

Table 3.1: Specified prior distributions for the three Bayesian conditions

Parameter	BayesDefault	BayesInfoI	BayesInfoII
Factor loadings	N(0, 0.01)	N(pop, 1)	N(pop, 1)
Regression coefficient β	N(0, 0.01)	N(0, 0.01)	N(pop, 1)
Variances latent variables*	G(1, 0.5)	G(1, 0.5)	G(1, 0.5)
Intercepts observed variables	N(0, 0.01)	N(0,0.01)	N(0, 0.01)
Residual variances observed variables*	G(1, 0.5)	G(1, 0.5)	G(1, 0.5)

Note. The column BayesDefault shows the `blavaan` default priors (Merkle & Rosseel, 2018). *Note that in `blavaan` the default priors are placed on precisions, which is the inverse of the variances. Abbreviations: N = Normal distribution with mean μ and precision τ; G = Gamma with shape α and rate β parameters on the precision (which equals an Inverse Gamma prior with shape α and rate β parameters on the variance); pop = population value used in data generation.

combination with `blavaan` default priors for the remaining parameters. Weakly informative priors were specified as follows: we set the mean hyperparameter of the normal distribution equal to the population value, and the precision hyperparameter equal to 1.

3.2.4 Evaluation Criteria

For each of the estimation methods and sample sizes, the occurrence of convergence problems and warnings will be assessed. For the parameter of interest, regression coefficient β, the following evaluation criteria will be used to evaluate the performance under the varying estimation methods and sample sizes: relative mean bias, relative median bias, mean squared error, coverage and power. All evaluation criteria will be computed across completed replications.[3]

Relative mean bias shows the difference between the average estimate across completed replications and the population value, relative to the population value. Relative median bias shows the relative difference between the median across completed replications and the population value.

[3]We defined completed replications as replications for which (1) the model did converge according to the optimizer and (2) for which for all parameters standard errors could be computed. If the model did not converge or standard errors were not computed for one or more parameters, we defined the replication as incomplete and excluded the replication from the aggregation of the results. All simulation code can be found in supplemental file S4 (https://osf.io/bam2v/).

The relative mean and median bias are computed by:

$$Relative\ mean\ bias = [^{(\overline{\theta}-\theta)}/_{\theta}] \times 100, \tag{3.1}$$

$$Relative\ median\ bias = [^{(\tilde{\theta}-\theta)}/_{\theta}] \times 100, \tag{3.2}$$

where $\overline{\theta}$ denotes the mean across completed replications, θ is the population value used for data generation, and $\tilde{\theta}$ denotes the median across completed replications. Values of relative mean and median bias below -10% or above $+10\%$ represent problematic levels of bias (Hoogland & Boomsma, 1998).

Mean squared error (MSE) is a combination of variability and bias across completed replications, where lower values indicate more stable and less biased estimates across replications. The MSE is computed by:

$$MSE = (\sigma)^2 + (\overline{\theta} - \theta)^2, \tag{3.3}$$

where σ is the standard deviation across completed replications, $\overline{\theta}$ denotes the average estimate across completed replications, and θ is the population value (Casella & Berger, 2002). A narrower distribution of estimates across replications (i.e., less variable estimates) leads to a smaller standard deviation across completed replications. Besides, the closer the estimated values are to the population value across completed replications, the smaller the amount of bias. MSE will be lower (and thus preferable) when the standard deviation and amount of bias across completed replications are small.

Coverage shows the proportion of completed replications for which the symmetric 95% confidence (for frequentist methods) or credibility (for Bayesian methods) interval contains the specified population value. Coverage values can range between 0 and 100, and values within the $[92.5; 97.5]$ interval are considered to represent good parameter coverage (Bradley, 1978).

Finally, statistical power is expressed as the proportion of estimates for which the 95% confidence (for frequentist methods) or credibility (for Bayesian methods) interval did not contain zero, across completed replications. Power values can range from 0 to 100, where values above 80 are preferred (Casella & Berger, 2002).

3.3 Results

3.3.1 Convergence

With small samples, we encountered severe convergence problems when frequentist methods were used, see Table 3.2. Differences between the three frequentist methods were especially visible when $n < 100$. With $n < 100$, twostep resulted in most non-converged cases, followed by ML, and finally followed by FSR.

The three Bayesian conditions produced results in all 1000 requested replications under all sample sizes.[4] However, when visually examining trace plots (for 2 randomly selected data sets x 6 sample sizes x 3 Bayesian conditions = 36 cases), severe convergence problems were detected for the smaller sample sizes, such as mode-switching, see Figure 3.2A. Mode-switching is defined as a chain that moves back and forth between different modes (Erosheva & Curtis, 2011; Loken, 2005), such as the chains in Figure 3.2A which move back and forth between values 5 and -5.

To further examine the extent of Bayesian convergence problems, we assessed trace plots for another 25 randomly selected data sets (resulting in 25 data sets x 6 sample sizes x 3 Bayesian conditions = 450 cases). In the assessment of these 25 selected data sets, mode-switching only occurred when BayesDefault was used when $n = 10$ or 20. Mode-switching disappeared when weakly informative priors were specified, see Figures 3.2B and 3.2C. Besides mode-switching, mild spikes were also detected when $n < 100$, see Figure 3.2D. Spikes are extreme values that are sampled during MCMC iterations, and could be seen as severe outliers. The appearance of spikes was reduced by the specification of weakly informative priors, see Figures 3.2E and 3.2F. From $n = 100$ onward, no convergence problems were detected when default priors were used. For more details on the convergence checks and more examples of trace plots, see supplemental file S2 (https://osf.io/bam2v/).

[4]Note that the number of iterations in the Bayesian analyses was automatically increased until the PSR criterion of 1.01 was reached.

Table 3.2: Number of completed replications, number of warnings about negative variance estimates, and number of completed replications without negative variance estimates for twostep, FSR and ML under varying sample sizes

	Completed replications[a]			Number (%) of warnings[b]			No negative variances[c]		
n	twostep	FSR	ML	twostep	FSR	ML	twostep	FSR	ML
10	475	641	533	259 (54.5%)	432 (67.4%)	446 (83.7%)	216	209	87
20	605	797	744	167 (27.6%)	360 (45.2%)	419 (56.3%)	438	437	325
50	809	970	955	41 (5.1%)	202 (20.8%)	217 (22.7%)	768	768	738
100	950	999	997	9 (0.9%)	58 (5.8%)	52 (5.2%)	941	941	945
250	1000	1000	1000	0	0	1 (0.1%)	1000	1000	999
500	999	1000	1000	0	1 (0.1%)	0	999	999	1000

Note. [a] Completed replications out of 1000 requested replications; [b] Number (%) of warnings of the completed replications; [c] Number of completed replications without negative variance estimates; n = sample size, twostep = twostep modeling; FSR = factor score regression, ML = maximum likelihood estimation

3.3.2 Warnings

For all small sample sizes, the three frequentist methods lead to a high percentage of warnings within the number of completed replications, see Table 3.2. All warnings were about negative variance parameters.[5] Differences between the three methods were especially present when $n < 100$. For these sample sizes, ML lead to the highest percentage of warnings, followed by FSR, and followed by twostep. As can be seen in Table 3.2, the number of warnings decreased when sample size increased. The number of completed replications without warnings about negative variance estimates is higher for twostep and FSR compared to ML, especially when $n < 100$.

For BayesDefault, three warnings about a small effective sample size occurred for $n = 10$, and two for $n = 20$.[6] No warnings occurred in the BayesInfoI and BayesInfoII conditions.

3.3.3 Results for Regression Coefficient β

In Figure 3.3, the relative mean bias (top) and relative median bias (bottom) are presented for the varying sample sizes and estimation methods. Because of the large discrepancy between the mean relative bias and median relative bias for sample sizes below 100, we plotted the complete distribution of parameter estimates for β across replications, see Figure 3.4. For all estimation methods, an increase in sample size led to: a decrease in the number of outliers; a narrower distribution of estimates (i.e., estimates are more stable across replications); and estimates closer to the population value. With samples as small as 10 and 20, the distributions of estimates are wider and a lot of outliers are present, which are signs of unstable estimates across replications. ML produced the most extreme outliers (up to 37.57 when $n = 10$). FSR and twostep show the narrowest distribution of estimates, indicating relatively stable behavior across replications. Overall, BayesInfoII

[5] The warning message that occurred for twostep, FSR and ML was: "some estimated ov [observed variables] variances are negative". For twostep and ML, a second message also occurred: "some estimated lv [latent variables] variances are negative".

[6] The warning message for BayesDefault: "Small effective sample sizes (< 100) for some parameters". The effective sample size expresses the amount of information in a chain while taking autocorrelation into account, for a more detailed explanation see Veen & Egberts (2020).

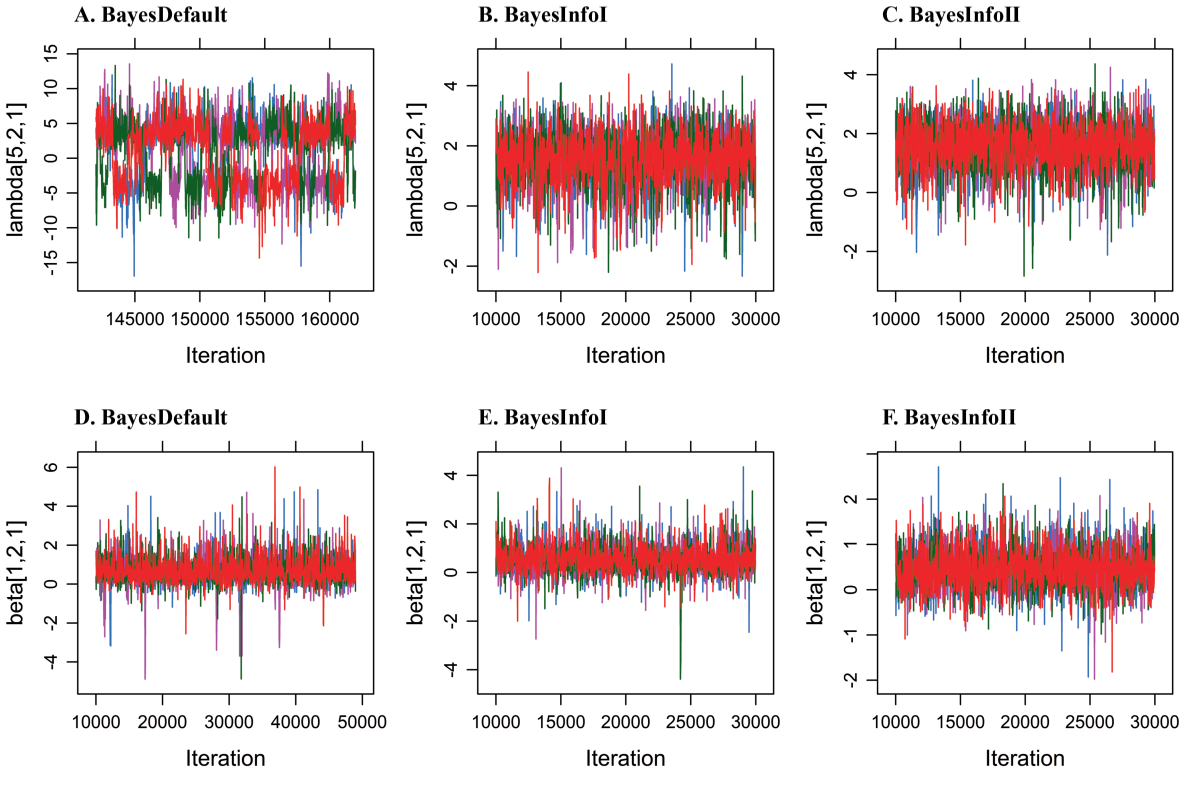


Figure 3.2: Trace plots for factor loading 5 (A–C), and regression coefficient β (D–F) after the analysis of BayesDefault, BayesInfoI and BayesInfoII. *Note.* Trace plots A–C correspond to the analysis of replicated data set 802 (within the simulation study) with a sample size of 10. Trace plots D–F correspond to the analysis of replicated data set 260 (within the simulation study) with a sample size of 20

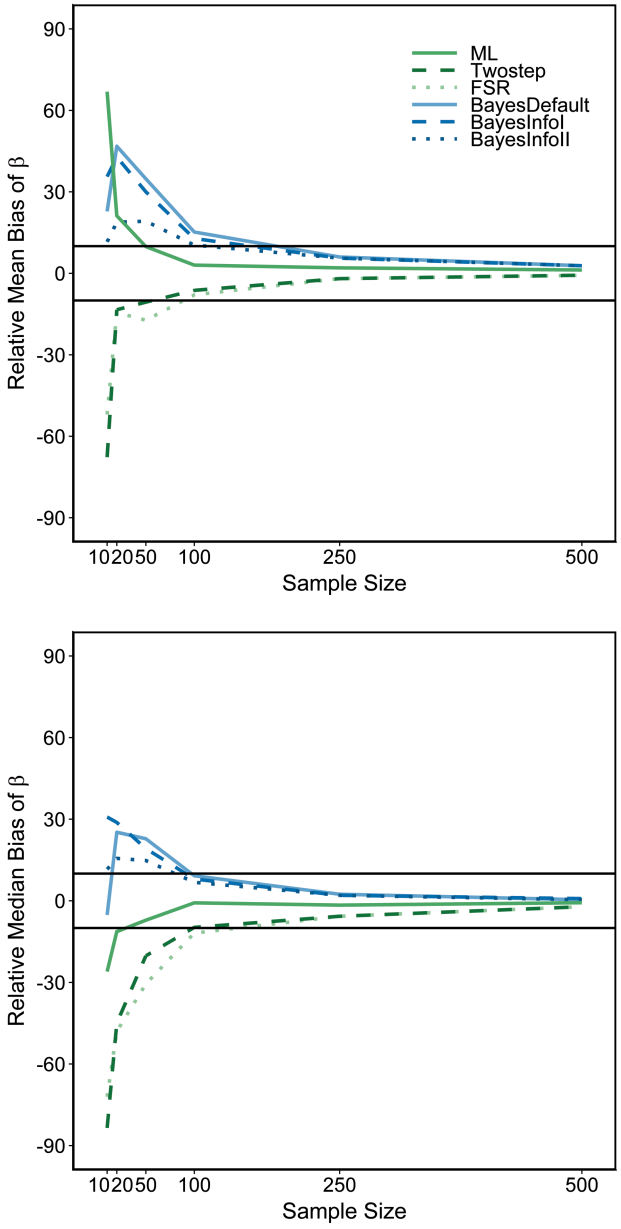


Figure 3.3: Relative Mean Bias (top) and Relative Median Bias (bottom) for parameter β, under varying sample sizes and estimation methods. *Note.* The static black horizontal lines represent the desired $\pm$ 10 % interval.

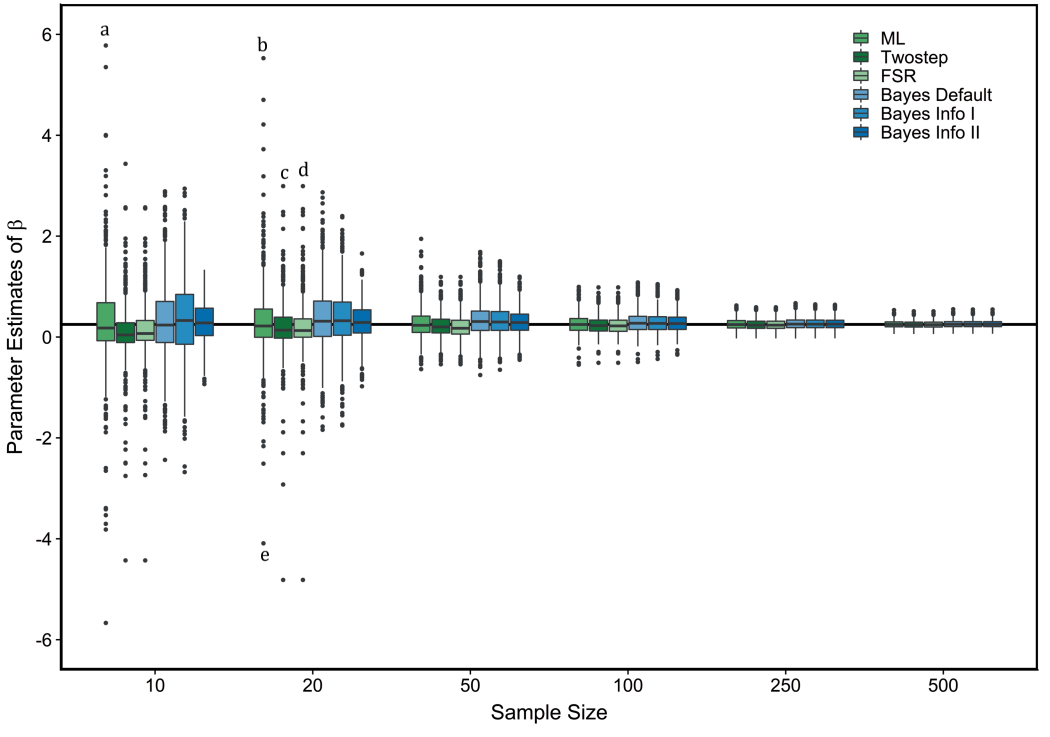


Figure 3.4: Distribution of the estimates for parameter β across completed replications, per estimation method and sample size. *Note.* The static black horizontal line denotes the true population value of 0.25 for β. Outliers are displayed as black circles, and outliers outside the interval [-6; 6] are denoted as follows: *a* denotes 11.39, 11.46, 14.87, 37.57 for ML when n = 10; *b* denotes 6.49, 8.89, 9.12 for ML when n = 20; *c* denotes 6.86, 6.89 for two-step when n = 20; *d* denotes 6.86, 6.89 for FSR when n = 20; and *e* -17.76 for ML when n = 20.

offers the best compromise between bias and stability: a narrow distribution of estimates, a mean and median close to the population value, and the smallest number of outliers. When $n = 100$, the differences between estimation methods become smaller; and the estimates become more stable across replications. For sample sizes of 250 and 500, differences between estimation methods are negligible and all estimation methods led to unbiased relative means and medians.

MSE for the regression coefficient β can be found in Figure 3.5A. Results are comparable to those shown in Figures 3.3 and 3.4. Differences between methods are especially visible when sample sizes are below 100. From $n = 100$ onward, MSE values are all close to zero. ML shows the highest MSE values for $n = 10$ and 20. BayesInfoI shows higher MSE than BayesDefault for $n = 10$, which was also visible in Figure 3.4 from the wider distribution of BayesInfoI relative to the distribution of BayesDefault for $n = 10$. The lowest MSE values are reported for BayesInfoII, followed by FSR, twostep, BayesDefault and BayesInfoI at $n = 10$. MSE values for FSR, twostep, BayesDefault and BayesInfoI are similar at $n = 20$, while BayesInfoII keeps the lowest MSE value. When $n = 50$ MSE values are comparable between methods, and from $n = 100$ onward the differences in MSE between methods are negligible.

Coverage results for regression coefficient β can be found in Figure 3.5B. All estimation methods show adequate coverage levels from $n = 100$ onward. For $n < 100$, the three Bayesian conditions show excessive coverage (> 97.50), although this slightly improved under BayesInfoI and BayesInfoII. Within the three frequentist methods, twostep and FSR resulted in higher coverage levels than ML. When $n < 100$, ML shows undercoverage (< 92.50), while FSR only shows slightly undercoverage when $n = 10$, and twostep when $n = 10$ and 20.

Results in terms of power can be found in Figure 3.5C. For all estimation methods, power is extremely low when the sample size is small, and only reached the desirable power level when $n = 500$. Across all sample sizes, the highest power levels are found for ML, followed by BayesInfoII, BayesInfoI, and twostep. The lowest power levels are found for FSR and BayesDefault.

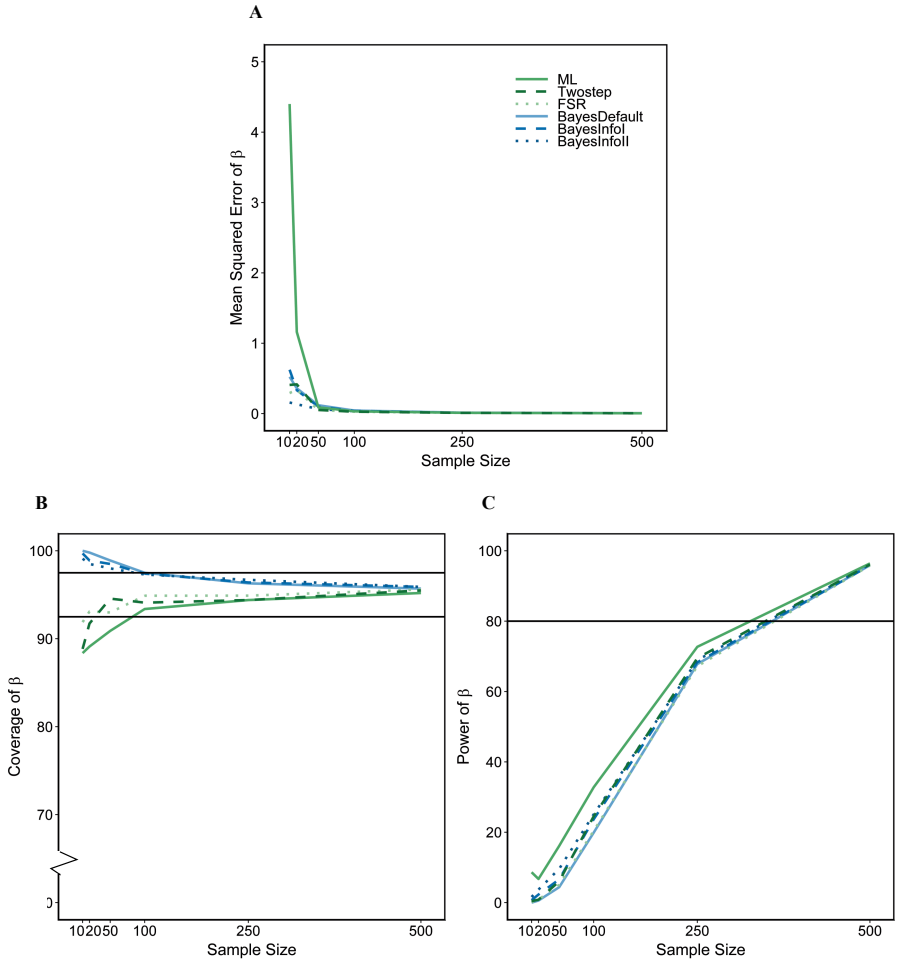


Figure 3.5: Mean Squared Error (A), Coverage (B), and Power (C) for parameter β, under varying sample sizes and estimation methods. *Note.* The static black horizontal lines in subfigure B represent the [92.5; 97.5] coverage interval, and the black horizontal line in subfigure C represents the desired 80% power level.

3.3.4 Results for Remaining Parameters

Besides regression coefficient β, 12 remaining parameters are estimated in the model: 2 variances for latent variables, 4 factor loadings and 6 residual variances.[7] In supplemental file S3 (https://osf.io/bam2v/), the distributions of parameter estimates across replications are displayed for the remaining parameters.

Estimates for these 12 parameters seem similar across estimation methods and have good statistical properties when $n = 250$ and 500. However, with sample sizes of 100 and below, frequentist methods show many (extreme) outliers and wide distributions, indicating unstable results across replications. Bayesian methods show notably fewer outliers and in general narrower distributions than the frequentist methods, especially under BayesInfoI and BayesInfoII conditions, although the medians of the distributions still deviate from the population values when $n < 100$.

3.4 Conclusion

In this chapter, we assessed – under varying sample sizes – the performance of three frequentist methods: twostep modeling (twostep), factor score regression (FSR) and maximum likelihood estimation (ML); and Bayesian estimation with three variations in prior specification. With sample sizes of 250 and 500, differences between estimation methods are negligible, and all methods led to stable and unbiased estimates. Consistent with existing simulation literature (e.g., Depaoli & Clifton, 2015; Hox & Maas, 2001; Van de Schoot et al., 2015) we found that ML led to severe convergence problems and a large amount of negative variance parameters when sample sizes are small. Compared to ML, both twostep and FSR led to better convergence rates without negative variances. Also, with small samples, twostep and FSR resulted in more stable results across replications and less extreme parameter estimates than ML. When Bayesian estimation was used with default priors, problematic mode-switching behavior of the chains did occur under small samples ($n = 10, 20$), even though the PSR values indicated that the overall model had converged.

[7] Note that when FSR is used, only three parameters are estimated: regression coefficient β, the variance of latent variable X and the variance of latent variable Y.

The presence of mode-switching can be a sign that the model is too complex for the data (Erosheva & Curtis, 2011).

Power is low for all estimation methods and only with a sample size of 500, the desired level of 80 was reached. The use of *weakly informative* priors (i.e., BayesInfoI and BayesInfoII conditions) instead of highly informative priors, as well as the specification of `blavaan` default priors for the remaining parameters, could explain why ML led to slightly higher power levels than Bayesian estimation in the current chapter (as opposed to previous studies, e.g., Miočević et al., 2017; Van de Schoot et al., 2015).

Also, the differences in power between default and informative prior conditions were smaller in the current chapter than expected. In previous studies (e.g., Van de Schoot et al., 2015; Zondervan-Zwijnenburg et al., 2019), priors with varying precision hyperparameters (e.g., 10 and 1) were compared to M*plus* default priors with a precision hyperparameter of 10^{-10} (Muthén & Muthén, 1998-2017). In the current chapter, the difference in precision hyperparameters between the informative ($precision = 1$) and default ($precision = 0.01$) conditions is noticeably smaller. This could explain why the increase in power with informative priors is lower in the current chapter than expected based on previous studies. Note that the level of informativeness of a prior distribution can only be interpreted relative to the observed data characteristics, and is therefore not generalizable to other studies (i.e., a weakly informative prior in one study can act as a highly informative prior in another study that uses different measurement instruments).

In summary, with extremely small sample sizes, all frequentist estimation methods showed signs of breaking down (in terms of non-convergence, negative variances, and extreme parameter estimates), as well as the Bayesian condition with default priors (in terms of mode-switching behavior). When increasing the sample size is not an option, we recommend using Bayesian estimation with informative priors. However, note, that the influence of the prior on the posterior is extremely large with relatively small samples. Even with thoughtful choices of prior distributions, results should be interpreted with caution (see also Chapter 4) and a sensitivity analysis should be performed, see Depaoli & Van de Schoot (2017) and Van Erp et al. (2018) on how to perform a sensitivity analysis. When no prior information

is available or researchers prefer not to use Bayesian methods, twostep and FSR are a safer choice than ML, although they can still result in non-convergence, negative variances, and biased estimates.

Note, however, that by adjusting the implementation of twostep and FSR, non-convergence problems could be circumvented by using an alternative non-iterative estimation method (instead of ML) to estimate the measurement and structural models (see Takane & Hwang, 2018); and as discussed in chapter 16 (Rosseel, 2020). In addition, negative variances could be avoided by restricting the parameter space to only allow positive values for variance parameters. Therefore, the preferred approach to implement twostep and FSR in small sample contexts should be further examined. We hope the current chapter is a starting point for future research in those directions.

Chapter 4

Dangers of the Defaults: A Tutorial on the Impact of Default Priors with Small Samples

This chapter is published as 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* [Special Issue on Quantitative Psychology and Measurement] https://doi.org/10.3389/fpsyg.2020.611963

Author Contributions: SS designed the tutorial manuscript and shiny app, and further developed the idea of the shiny app with SW. SW worked out the code for the shiny app with input and feedback from SS. SS took the lead in writing the manuscript. SW wrote the "Shiny App" section and provided feedback on the manuscript.

Online Data Archive and Supplementary Files: https://osf.io/m6byv/

Abstract

When Bayesian estimation is used to analyze Structural Equation Models (SEMs), prior distributions need to be specified for all parameters in the model. Many popular software programs offer default prior distributions, which is helpful for novel users and makes Bayesian SEM accessible for a broad audience. However, when the sample size is small, those prior distributions are not always suitable and can lead to untrustworthy results. In this chapter, we provide a non-technical discussion of the risks associated with the use of default priors in small sample contexts. We discuss how default priors can unintentionally behave as highly informative priors when samples are small. Also, we demonstrate an online educational Shiny app, in which users can explore the impact of varying prior distributions and sample sizes on model results. We discuss how the Shiny app can be used in teaching; provide a reading list with literature on how to specify suitable prior distributions; and discuss guidelines on how to recognize (mis)behaving priors.

4.1 Introduction

Bayesian estimation of Structural Equation Models (SEMs) has gained popularity in the last decades (e.g., Kruschke et al., 2012; Van de Schoot, Winter, et al., 2017), and is more and more often used as a solution to problems caused by small sample sizes (e.g., König & Van de Schoot, 2017; McNeish, 2016a).[1] With small samples, frequentist estimation (such as [restricted] Maximum Likelihood or [weighted] least squares estimation) of SEMs can result in non-convergence of the model, which means that the estimator was unable to find the maximum (or minimum) for the derivative of the model parameters. Even when a model converges, simulation studies have shown that the parameter estimates may be inadmissible (e.g., Heywood cases) or inaccurate (i.e., the estimate deviates from the population value; Boomsma (1985); Nevitt & Hancock (2004)). In contrast to frequentist methods, Bayesian methods do not rely on large sample techniques, which make Bayesian methods an appealing option when only a small sample is available. Within the Bayesian framework, prior distributions need to be specified for all parameters in the model.[2] This additional step may pose a barrier for novice users of Bayesian methods. To make Bayesian SEM accessible to a broad audience, popular software programs for analyzing Bayesian SEMs, such as M*plus* (Muthén & Muthén, 1998-2017) and the `blavaan` package (Merkle & Rosseel, 2018) in R (R Core Team, 2022), offer default prior distributions. However, those default prior distributions are not suitable in all cases. When samples are small, the use of solely default priors can result in inaccurate estimates—particularly severely inaccurate variance parameters—unstable results, and a high degree of uncertainty in the posterior distributions (e.g., Gelman, 2006; McNeish, 2016a; Smid, McNeish, et al., 2020). These three consequences of using default priors with small

[1] There are many other reasons why researchers use Bayesian SEM, such as the ability to estimate models that are not identified in the frequentist framework or to resolve issues with missing data, non-linearity, and non-normality (see e.g., Kaplan, 2014, pp. 287–290; Van de Schoot, Winter, et al., 2017; Wagenmakers et al., 2008). However, the focus of this chapter is the use of Bayesian estimation to deal with small samples.

[2] Prior distributions represent information about the parameters and can be based on previous studies or the beliefs of experts in the field. The prior distributions are then updated by the likelihood (observed data depended on the model). By using methods such as Markov chain Monte Carlo (MCMC), the posterior distribution is simulated, which is a combination of the prior and likelihood. For references with an elaborate introduction into Bayesian estimation, we refer to our reading list (https://osf.io/pnmde).

samples severely limit the inferences that can be drawn about the parameters in the model.

With small samples, the performance of Bayesian estimation highly depends on the prior distributions, whether they are software defaults or specified by the researcher (e.g., Gelman et al., 2014; Kaplan, 2014; McElreath, 2016). McNeish (2016a) discussed that small sample problems (such as non-convergence, inadmissible and inaccurate parameter estimates) cannot be fixed by only switching from a frequentist to a Bayesian estimator. Instead, he argues that if Bayesian methods are used with small samples, "prior distributions must be carefully considered" (McNeish, 2016a, p. 764). This advice is not new: Kass & Wasserman (1996) already warned against relying on default prior settings with small samples. In the quarter-century since that initial warning, Bayesian estimation is increasingly used to deal with small samples (Smid, McNeish, et al., 2020; Van de Schoot, Winter, et al., 2017). Yet researchers remain stubbornly reliant on default priors, despite clear caution against their use (as shown by König & Van de Schoot, 2017; McNeish, 2016a; Van de Schoot, Winter, et al., 2017).

4.1.1 Goals of this Tutorial

In this chapter, we provide a non-technical discussion of the risks associated with the use of default priors. We discuss how default priors can unintentionally behave as highly informative priors when samples are small. Next, we demonstrate an educational online Shiny app (Smid & Winter, 2020, available on our Open Science Framework (OSF) page via https://osf.io/m6byv), in which users can examine the impact of varying prior distributions and sample size on model results. We discuss how the Shiny app can be used in teaching and provide an online reading list (available via https://osf.io/pnmde) with literature on Bayesian estimation, and particularly on how to specify suitable prior distributions. Finally, we provide guidelines on how to recognize (mis)behaving priors.

4.2 What is a Small Sample?

Before we continue our discussion of the potential dangers of default priors with small samples, we need to address the question: What exactly is a small sample? Whether a sample is small depends on the complexity of the model that is estimated. One way to express the size of a sample is to look at the ratio between the number of observations and the number of unknown parameters in the model (e.g., Lee & Song, 2004). A sample could be considered very small when this ratio is 2, which means there are just two observations for each unknown parameter. As SEMs often include many unknown parameters (i.e., factor loadings, intercepts, covariances), samples that may appear relatively large are in fact very small. For example, a confirmatory factor analysis model with three latent factors and fifteen observed items consists of 48 unknown parameters: 12 factor loadings (first factor loading fixed at 1 for identification), 15 intercepts, 15 residual variances, 3 factor variances, and 3 factor covariances. In this scenario, a sample of 100 participants would still be considered very small (ratio = 2.08). This example demonstrates that general rules of thumb about sample sizes for SEM (e.g. $n > 100$, see Kline, 2015) can be misleading as they do not take into account model complexity. Furthermore, model complexity depends on more than just the number of parameters that are estimated. Other factors that play are role in model complexity are whether the model includes components such as categorical variables, latent factors, multiple groups, or latent classes. A review of simulation studies on SEM (Smid, McNeish, et al., 2020, see chapter 1 of this dissertation) showed that authors of these simulation papers have widely varying definitions of a "small sample size", ranging from extremely small (e.g., n = 8 assessed at three time points with one continuous variable, see Van de Schoot et al., 2015) to what some might consider moderately sized (e.g., n = 200 with 12 ordinal variables, see Chen et al., 2015). Thus, assessing whether a sample is (too) small is unfortunately not as easy as checking whether a certain number of participants has been reached, and should be done on an analysis-by-analysis basis.

4.3 Dangers of the Defaults

The risks associated with default priors when Bayesian SEM is used with small samples can be described as a combination of the following three factors.

First, when samples are small, priors have a relatively larger impact on the posterior than when samples are large. The posterior can be seen as a compromise between the prior and the likelihood. With a larger sample size, the likelihood dominates the posterior (see Figure 4.1C). However, with a small sample size, the likelihood has relatively less weight on the posterior. Accordingly, the prior has relatively more weight on the posterior (see Figure 4.1A). Therefore, it is of great importance to specify suitable prior distributions when samples are small (e.g., Gelman et al., 2014).

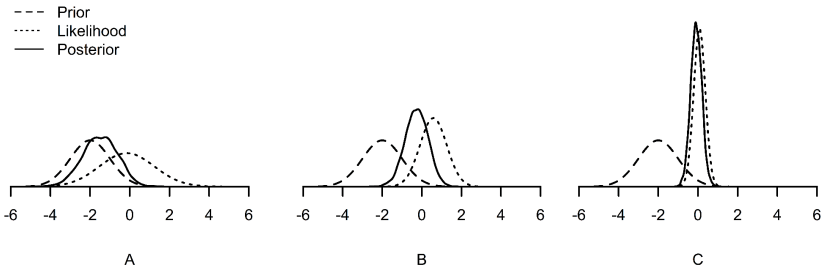


Figure 4.1: Examples of prior, likelihood and posterior distributions under small (A), medium (B), and large (C) sample sizes. The posterior distribution is dominated by the prior under the small sample size (A), and dominated by the likelihood under the large sample size (C).

Second, most of the default priors have very wide distributions. For instance, the M*plus* default prior for means and regression coefficients is a Normal distribution with a mean hyperparameter of zero and a variance of 10^{10} (Muthén & Muthén, 1998-2017). The variance hyperparameter corresponds to a standard deviation of 100.000, meaning, that 68% of the prior distribution contains values between -100.000 and 100.000, and 95% of the prior distribution contains values between -200.000 and 200.000.[3] When such

[3]Hyperparameters are the parameters of prior distributions, such as the mean and variance of the Normal distribution, and the alpha and beta in inverse gamma.

default priors are specified, a wide range of parameter values can be sampled from the posterior during the Bayesian analysis. All those parameter values are therefor considered plausible, which might not always be appropriate. For instance, when measuring mathematical ability on a scale from 0 to 100, values below 0 and above 100 cannot be present in the data. Specifying a default prior with such a wide distribution on the mean of mathematical ability will put a lot of weight on values that are not reasonable (see e.g., Stan Development Team, 2017, p. 131). For small sample sizes, the combination of the relatively larger impact of the prior on the posterior and the wide distribution of default priors can lead to extremely incorrect parameter estimates (see e.g., Gelman, 2006; McNeish, 2016a; and the systematic literature review of Smid, McNeish, et al., 2020).

The third factor that plays a role, is the false belief that default priors are noninformative priors which 'let the data speak'. Default priors can act as highly informative priors, as they can heavily influence the posterior distribution and impact the conclusions of a study (see e.g., Betancourt, 2017). As explained by McNeish (2016a) (p. 752): "with small samples, the idea of noninformative priors is more myth than reality (...)". The terminology of informative and noninformative priors can therefore be confusing (see also Bainter, 2017, p. 596). In addition, different software programs use different default priors (see Table 4.1).

Van Erp et al. (2018, p. 26) investigated the performance of multiple default priors and concluded that, especially with small samples, all investigated default priors performed very differently, and "that there is not one default prior that performed consistently better than the other priors (...)." The choice of software could thus unintentionally influence the results of a study (see e.g., Holtmann et al., 2016), which is problematic if one is not aware of this. Note that we are not advocating against default priors in general.

Default priors can be suitable — even when samples are small — in cases where all values in the prior distribution are reasonable and can occur in the data (for example values around 100,000 or 200,000 are realistic in housing price data, see e.g., LeGower & Walsh, 2017). However, the use of default priors is problematic when researchers assume they let 'the data speak' while in reality they 'let the default priors speak', meaning that the priors can heavily impact the results without one being aware of this.

Table 4.1: Overview of default prior distributions in the software program M*plus* and the R package `blavaan`

	M*plus* priors (v. 8.4) on variance σ^2	`blavaan` priors (v. 0.3-8) on precision $1/\sigma^2$ or standard deviation σ
Observed variable intercept	N(0, 10^{10})	N(0, 32)
Latent variable intercept, factor loading, regression	N(0, 10^{10})	N(0, 10)
Variance covariance blocks of size 1	IG(-1, 0)	
Variance covariance blocks of size > 1	IW(0, -p, -1), where p is the size of the matrix	
Observed and latent variable variance		G(1, 0.5)[a]
Covariance matrix		W(3, I)[b]
Correlation		B(1, 1)
Threshold	N(0, 10^{10})	N(0, 3.16)

Note. Prior distributions in M*plus* are placed on the variance, while the prior distributions in `blavaan` are placed on the precisions (the inverse of the variance) unless stated otherwise. N = Normal distribution with hyperparameters mean μ and variance σ^2; I = Identity Matrix; IG = Inverse Gamma; G = Gamma; IW = Inverse Wishart; W = Wishart; B = Beta distribution. [a] The prior for the observed and latent variable parameters is placed on the standard deviation. [b] In `blavaan`, three MCMC packages can be used (target "stan", "stanclassic" and "jags") for the analysis. For all the MCMC packages, the same default priors are specified, with one exception: for target = "jags", a different prior for the covariance is specified.

In the next section, we discuss the Shiny app that we developed to demonstrate in an example the possible informative behavior of default priors when the sample is small.

4.4 Shiny App: The Impact of Default Priors

We have created a Shiny app that serves as an educational tool that can be used to learn more about the impact of default priors in Bayesian SEM. It can be found online via https://osf.io/m6byv, together with supplemental files and R code to reproduce the app. In addition, we have created a lesson plan (available for download in the app) to support the educational focus of the app. The app consists of three pages: (1) a page where users can interactively explore the impact of prior settings and sample size on a Bayesian latent growth model (see Figure 4.2), (2) an overview of the prior specifications used in the app, and (3) a list of further resources to learn more about various aspects of Bayesian SEM. The main, interactive, page includes a menu that walks users through selecting their sample size, prior specification settings, and running the model a first time and a second time with a doubled number of iterations (in line with the WAMBS checklist of Depaoli & Van de Schoot, 2017). The models in the Shiny app were externally run using the software M*plus* (Muthén & Muthén, 1998-2017) to enhance the user experience.[4]

The main window on the page has five tabs that can be used to (1) see what model is estimated, (2) check convergence of the model using the potential scale reduction factor (PSFR, Gelman & Rubin, 1992), examine the precision of the posterior samples with the effective sample size (ESS), (3) look at plots of the prior, likelihood, and posterior and trace plots, (4) inspect parameter estimates, (5) access the lesson plan.

[4]This popular, user-friendly software program for estimating Bayesian SEM has made it extremely easy to be a naive user of Bayesian statistics (one only needs to include the line "Estimator = Bayes;" in the input file).

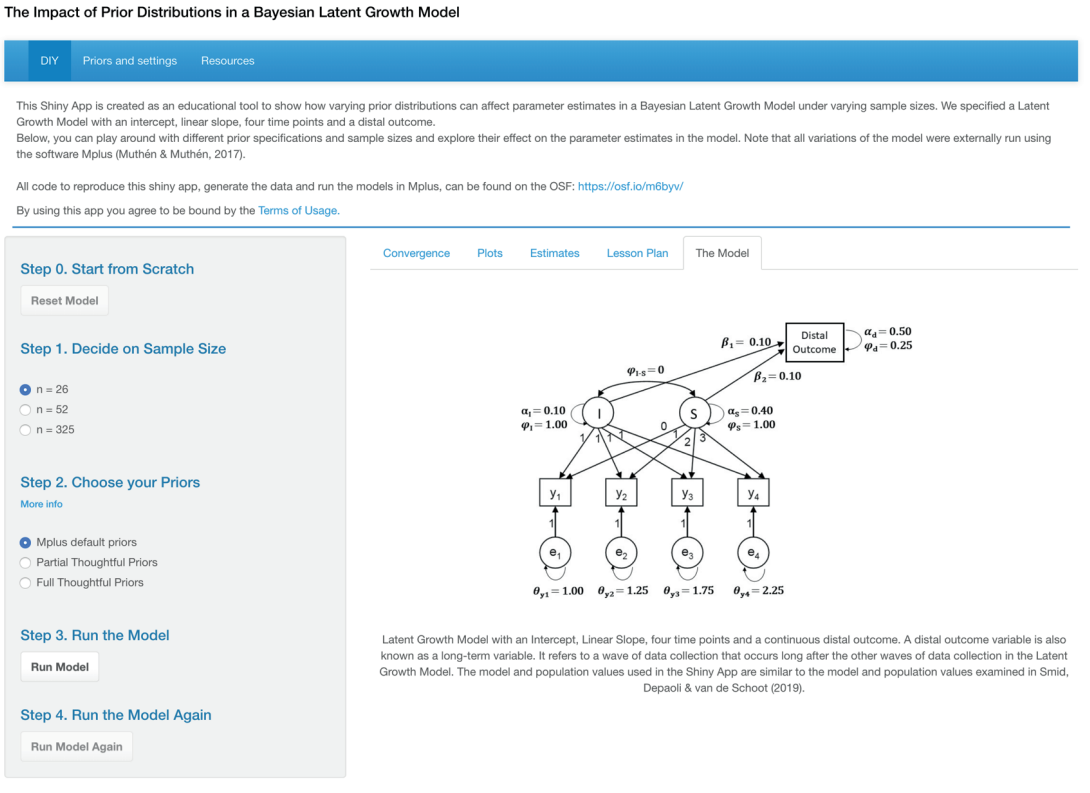


Figure 4.2: Main page of the Shiny app, where users can interactively explore the impact of prior settings and sample size in a Bayesian Latent Growth Model

4.4.1 The Model, Sample Sizes, and Priors used in the Shiny App

The model, sample sizes, and prior settings used in the Shiny app are based on Smid, Depaoli, & Van de Schoot (2020). Specifically, the model is a latent growth model (LGM) with a latent intercept and linear slope, four time points, and a continuous long-term variable (i.e., distal outcome) that is predicted by the latent intercept and slope (see Figure 4.3). A long-term variable is a variable that is collected at a wave of assessment that occurs long after the other waves of assessment in the LGM. An example of a distal outcome is young adult levels of depression that are predicted by conduct and emotional problems at ages 4 to 16 (Koukounari et al., 2017). Users can select one of three sample sizes: 26, 52, 325, which represent a very small, small, and relatively large sample for the model of interest, which has 13 unknown parameters.

Three different prior specifications are included in the app: one specification using software default priors and two specifications with increasing numbers of thoughtful priors. The default priors that we selected are those specified in M*plus* (Muthén & Muthén, 1998-2017) and are called "M*plus* default priors" in the Shiny app. The two thoughtful prior specifications, called "Partial Thoughtful Priors" and "Full Thoughtful Priors", were taken from Smid, Depaoli, et al. (2020), details of which are included on the second page of the Shiny app. In short, "Partial Thoughtful Priors" includes informative priors for the mean of the intercept and slope of the LGM, the regression coefficients, and the intercept of the distal outcome. "Full Thoughtful Priors" includes informative priors on all parameters in the model, with the exception of the residual variances. These two specifications reflect scenarios where a researcher has access to prior knowledge regarding some or most of the parameters in the model.

The specific hyperparameter values of the thoughtful priors (e.g., where the center of the prior is and how narrow the prior is) in the example used in the app are somewhat arbitrary because they are based on a simulation study. Specifically, the priors are all centered around the (known) population values and the width of the priors is based on the width of the posterior distribution of the analysis done with M*plus* default priors. This approach is most closely related to a type of prior specification called data dependent prior

specification (McNeish, 2016b), where an initial analysis using default priors or frequentist estimation methods provides the values for the prior hyperparameters. In applied research, data dependent priors are controversial, as the researcher technically double-dips by using their data to specify the priors that are subsequently used to analyze their data (Darnieder, 2011). To resolve this issue, researchers could split their data in half and base the prior specification for the Bayesian analysis on the results of a frequentist analysis using 50% of the total sample. As this approach would further reduce the sample size for the final analysis, this approach for specifying priors may not be feasible with small sample sizes.

The two thoughtful prior specifications included in the app are just two examples of how thoughtful priors can be included in Bayesian SEM. Other sources that can be used for specifying thoughtful priors include previous research, meta-analyses, or knowledge from experts in the field (for in-depth discussions of these topics, we refer to Lek & Van de Schoot, 2018; Van de Schoot et al., 2018; Zondervan-Zwijnenburg et al., 2017). Even if prior knowledge is not readily available, researchers can think about impossible and implausible values for the parameters and specify prior distributions that only contain information about the typical range of the parameters. To illustrate this idea, imagine that the distal outcome of the LGM shown in Figure 4.3 was measured with a questionnaire that had a range from 0 to 20. A researcher could use this information to specify a prior for the intercept of the distal outcome that makes values outside of that range highly improbable (e.g., N(10, 15)). For some parameters, it may be challenging to identify prior hyperparameters that will exclude implausible values. For example, the inverse Gamma distribution is often used as a prior for the (residual) variance parameters. The parameters of this distribution, called shape and scale, are not as easily interpreted and thoughtfully specified as the mean and variance of a normal distribution. Fortunately, methods for specifying thoughtful prior hyperparameters for the inverse Gamma distribution have been suggested (e.g., Zitzmann et al., 2021). Alternatively, researchers may decide to switch to a different distribution altogether (Van Erp et al., 2018). Examples include the half-Cauchy prior (Gelman, 2006; Polson & Scott, 2012) or reference priors such as Jeffrey's prior (Tsai & Hsiao, 2008).

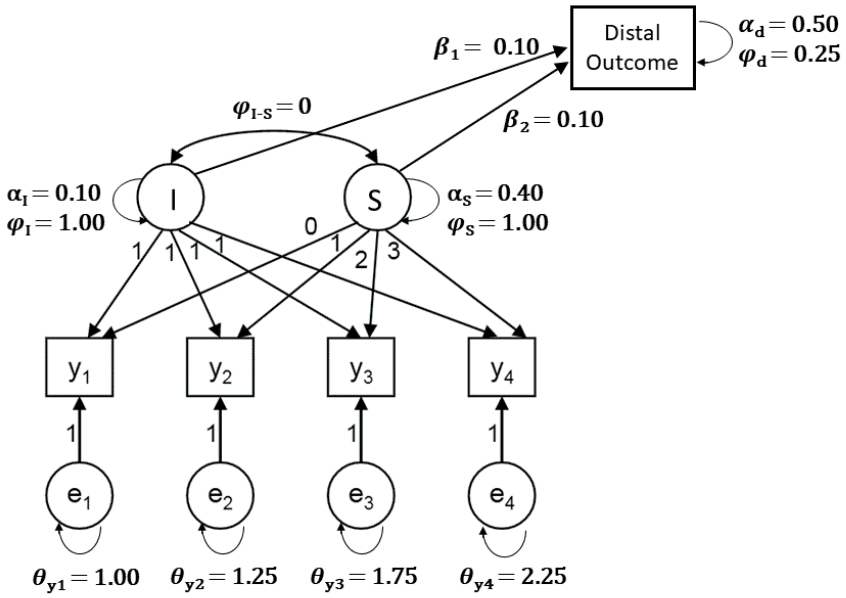


Figure 4.3: The Latent Growth Model with a distal outcome variable that is used in the Shiny app, including population values. Model and population values are based on Smid, Depaoli, and van de Schoot (2020).

4.4.2 Using the Shiny App as a Teacher

Since this Shiny app was explicitly developed to serve as an educational tool, we have created a worksheet and answer key that can be downloaded directly in the app itself.[5] In addition, it is possible within our app to export all plots and tables created. These can be used in answering the questions on the worksheet. By making students aware of the impact of relying on default settings when samples are small, we hope to teach students about the importance of specifying suitable prior distributions and to contribute to the responsible use of Bayesian SEM.

4.5 Guidelines: How to Recognize a (Mis)behaving Prior?

To formulate suitable prior distributions and to check afterwards whether the priors are 'behaving', information is needed about the reasonable range of values for the parameters in the model. This information can be based on previous studies, the scale or questionnaire that is used, or expert knowledge from the field. In our reading list (available via https://osf.io/pnmde), we provide an overview of relevant literature on how to specify suitable priors based on multiple sources of information. Below, we discuss four ways to identify a (mis)behaving prior after conducting a Bayesian analysis (see also Table 4.2), by inspecting for all parameters the (a) effective sample size, (b) trace plots, (c) prior-likelihood-posterior distributions, and (d) the posterior standard deviation and 95% highest posterior density.

4.5.1 Effective Sample Size

Inspecting the effective sample size (ESS) of each parameter in the model is a good first step in the search for misbehaving priors. The ESS represents the number of independent samples that have the same precision as the total number of samples in the posterior chains (Geyer, 1992). The ESS is closely related to the concept of autocorrelation, where current draws from the

[5] The worksheet can be found on the main page under the fifth tab ("Lesson Plan").

Table 4.2: Possible signs of 'misbehaving' priors

Effective sample size
- Low effective sample size (i.e., < 1,000) can be a first indication that the priors are problematic
Trace plots
- Spikes: shape of alien communication captured in a sci-fi movie instead of a fat caterpillar
- Highly improbable values for the parameter on the y-axis based on information about the reasonable range of values about parameters
- Chains that are not overlapping
Prior-likelihood-posterior comparison
- Substantial deviation between prior, likelihood and/or posterior: e.g., a posterior that is much narrower or wider than the prior and likelihood, while taking into account the amount of information in the prior (i.e., level of informativeness of the prior) and in the likelihood (i.e., sample size)
Posterior SD and 95% HPD
- Much smaller or larger posterior SD or 95% HPD than expected based on the amount of information in the prior (i.e., level of informativeness of the prior) and in the likelihood (i.e., sample size)

posterior distribution are dependent on previous draws from the posterior distribution. Autocorrelation is undesirable as it increases the uncertainty in posterior estimates. If autocorrelation within the chains is low, then the ESS approaches the total number of samples in the posterior chains, and the posterior distribution will be more precise and more likely to approximate the parameter estimate well (Zitzmann & Hecht, 2019). If autocorrelation within the chains is high, a larger number of samples will be necessary to reach an adequate ESS. A low ESS can be the first indicator that there might be a misbehaving prior. Multiple recommendations have been made about how to assess whether the ESS is too low: Zitzmann & Hecht (2019) recommend that ESSs should ideally be over 1,000 to ensure that there is enough precision in the chain. It is also possible to compute a lower bound for the number of effective samples required using a desired level of precision and the credible interval level of interest (Flegal et al., 2021; Vats, Flegal, & Jones, 2019). Finally, it can also be helpful to look at the ratio of the ESS to the total number of samples, where a ratio < 0.1 indicates that there are high levels of autocorrelation in the chains (although this does not necessarily indicate that the posterior distribution is not precise, see Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019). A low ESS can serve as the first clue that something might be wrong, but even if all ESSs appear

acceptable, plots and posterior estimates should be inspected to further confirm if priors are behaving.

4.5.2 Trace plots

Three characteristics of a trace plot can indicate a misbehaving prior. First, the shape of the trace plot: If the multiple chains are well-behaved, the chains should resemble the hungry caterpillar after six days of eating (see Figure 4.4A). A misbehaving prior can result in trace plots that exhibit spikes, closely resembling alien communication captured in a sci-fi movie (Figure 4.4C). Second, do the values that are covered by the posterior make sense for this parameter, or is the y-axis stretched to cover unrealistic values? Even when subtle spikes are present (Figure 4.4B), the y-axis range could show that the chains are drawing improbable values from the posterior distribution and should be given extra attention. Third, a lack of overlap of the chains can indicate a misbehaving prior. When the chains do not overlap, it indicates that they are sampling from different parts of the posterior distribution and are not converging towards the same location.

4.5.3 Prior-Likelihood-Posterior Comparison

One important aspect of our Shiny app is that the prior, likelihood, and posterior distributions are visualized to make comparisons across different priors and sample size settings easy.[6] When there is a substantial deviation between the prior, likelihood and posterior distributions, results should be interpreted with caution, especially when the sample size is small. Researchers should decide how much impact of the prior and likelihood on the posterior is desirable. Is it preferable that the posterior is a compromise between the prior and likelihood, or that the posterior is dominated by one of two? For instance, when the likelihood and the prior deviate a lot, one might not want to trust the posterior results.[7]. In case of small samples, the

[6]For details on how we visualized priors, likelihood and posterior distributions, we refer to the PLPPFunction.Mplus.R file on the OSF (https://osf.io/m6byv).

[7]For readers interested in the impact of so-called prior-data conflict, we refer to simulation studies by Depaoli (2014); Holtmann et al. (2016); and Smid, Depaoli, et al. (2020)

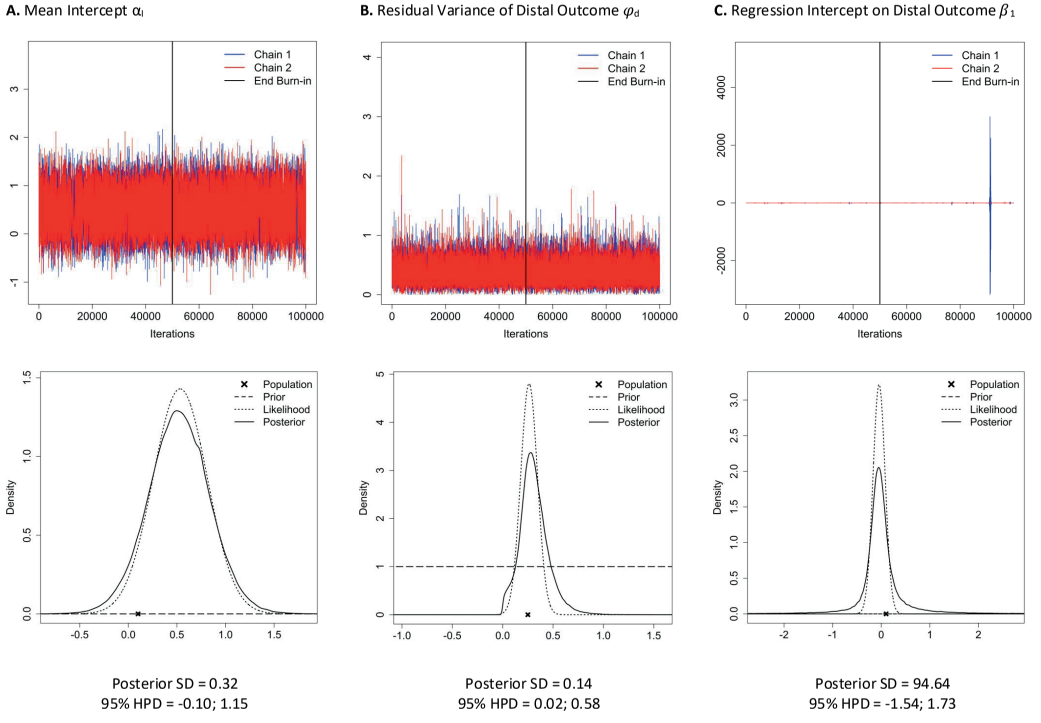


Figure 4.4: Traceplots; prior, likelihood, posterior plots; posterior standard deviation (SD) and 95% highest posterior density interval (HPD) for three parameters: mean intercept (A), residual variance of the distal outcome (B) and the regression effect of the slope on the distal outcome (C) under sample size n = 26 and M*plus* default priors (examples retrieved from the Shiny app). Note that the M*plus* default prior for residual variance parameters is IG(-1, 0), which is improper (i.e. does not integrate to 1) and has a constant density of 1 on the interval (-∞, ∞) (Asparouhov & Muthén, 2010).

results might especially be driven by the prior distributions. This is only desirable when researchers trust the specified prior distributions. Figure 4.4 shows the prior-likelihood-posterior comparison for three parameters. Although the prior distributions (dashed lines) look completely flat, default prior distributions were used for all parameters. In Figure 4.4A, the posterior (solid line) closely follows the likelihood distribution (dotted line), which is desirable here because the default prior (dashed line) is specified and we do not want it to impact the posterior much. In Figures 4.4B and 4.4C, the posteriors seem to have tails that are too fat (kurtotic) compared to the likelihood distribution and the flat default priors, and results should therefore be inspected further.

4.5.4 Posterior SD and 95% HPD

The posterior standard deviation (SD) and 95% credible (or highest posterior density; HPD) interval can be inspected to assess whether the estimates are unusually certain or uncertain. Uncertainty is demonstrated by a large posterior SD and a wide 95% HPD. Available information about reasonable values for the parameters as well as the amount of information in the prior and likelihood should be used to assess whether the level of (un)certainty of the posterior is reasonable. For instance, in Figure 4.4C, a posterior SD of 94.64 is reported, which is a much higher value than would be expected for a regression estimate and implies that some very extreme values were likely sampled from the posterior. This level of uncertainty is also reflected by the extreme spikes in the trace plot and the kurtotic posterior distribution. The parameters depicted in Figure 4.4 illustrate that the combination of a noninformative prior and a small sample size does not always lead to problems across all parameters in a model. It is important to note that even if it appears that the priors of the main parameter(s) of interest are behaving well, a misbehaving prior that is located elsewhere in the model may lead to inaccuracies in the posterior estimates of the main parameters. For example, in a multilevel SEM with a between-level covariate effect, the between-level variance estimate may not be of substantive interest. However, a supposedly noninformative prior [IG(.001,.001)] for the between-level variance parameter can turn into a misbehaving prior when the amount of variance located at the between-level is large (Depaoli & Clifton, 2015). In a simulation study,

Depaoli & Clifton (2015) showed that this misbehaving prior resulted in a biased posterior estimate of the between-level covariate effect. A researcher who only inspected the trace plot for the between-level covariate effect may not have realized that their results were negatively affected by a prior placed on between-level variance parameter. For that reason, it is critical to always examine all parameters in the SEM.

4.5.5 What to do if you suspect a misbehaving prior?

When one of the trace plots, prior-likelihood-posterior distribution plots, posterior SDs or 95% HPDs show signs of a misbehaving prior, results should not be trusted, and researchers should proceed with caution. Unfortunately, we cannot provide rules of thumb for when these indicators of misbehavior become problematic. It depends on the specified prior, the data, the parameter, the model of interest, and the personal judgement of the researcher. A sensitivity analysis can help assess the impact of the specified prior distributions on the posterior (see Depaoli & Van de Schoot, 2017; Van Erp et al., 2018). Again, it is up to the researcher to decide whether a certain amount of impact of the prior is desirable or not. Therefore, Bayesian SEM should only be used with small samples when researchers are able and willing to make these types of decisions.

4.5.6 Reporting of Bayesian SEM

Although a rich body of literature exists on good practice of how to perform and what to report for a Bayesian analysis (see e.g., Depaoli & Van de Schoot, 2017; Kruschke, 2015, pp. 721–725), we want to stress the importance of transparency and reporting every decision. We advise to always provide an (online) appendix in which is explained in detail which priors are specified and why these specific priors are chosen. For more literature and examples on reporting Bayesian SEM, we refer to our reading list on https://osf.io/pnmde.

4.6 An Illustration: The Impact of Default Priors

To illustrate the impact of prior settings and sample size—and the informative behavior of default priors with a small sample size—we retrieved the trace plots, prior-likelihood-posterior plots, and posterior SDs from the Shiny app for a single parameter: the regression effect of the distal outcome regressed on the linear slope ($\beta 2$ in Figure 4.3). The plots (Figure 4.5 show signs of a misbehaving prior when samples are small (n = 26, or 52 for this model) when default priors are used. Specifically, the trace plots exhibit spikes that reach highly improbable values for the regression coefficient, the plots have a stretched y-axis, and show chains that are not overlapping. Moreover, the prior-likelihood-posterior plots for the two small sample sizes show that the posterior distribution (solid line) is wider than the likelihood estimate (dotted line). Overall, the plots displayed in Figures 4.5 show that default priors, which are assumed to be noninformative, can impact the results when samples are small. Options for improving model estimation include increasing the sample size or specifying suitable priors for the parameters.

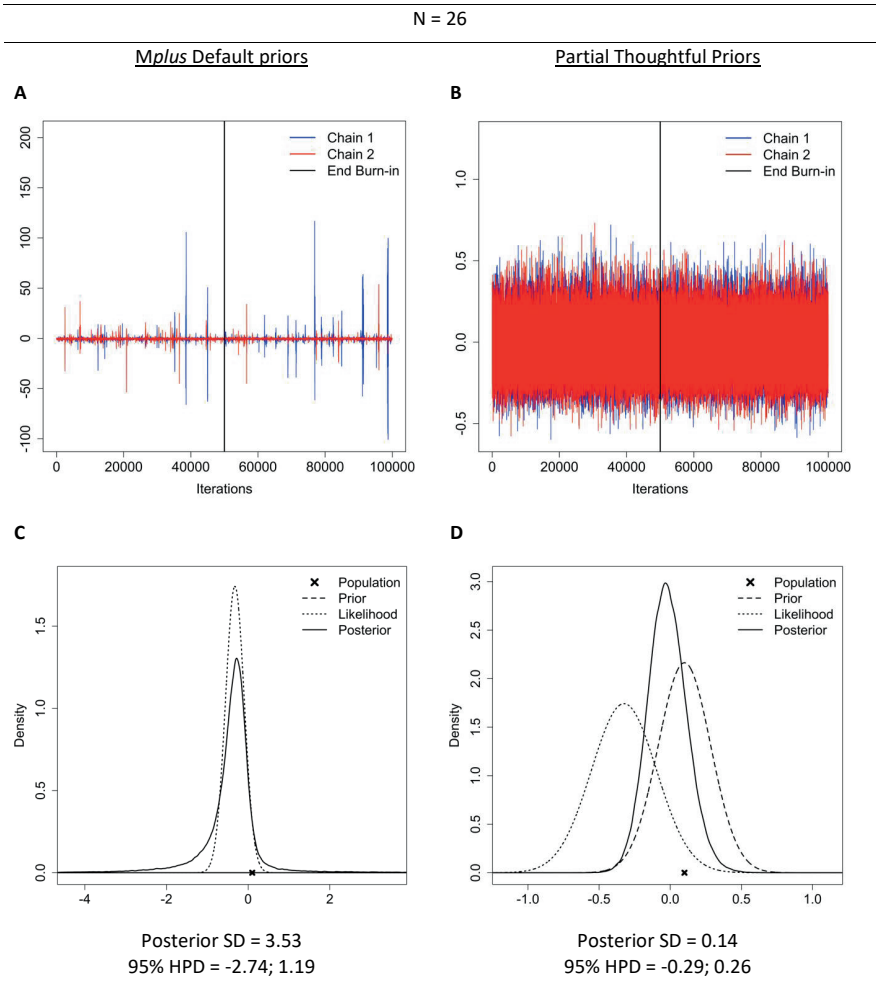
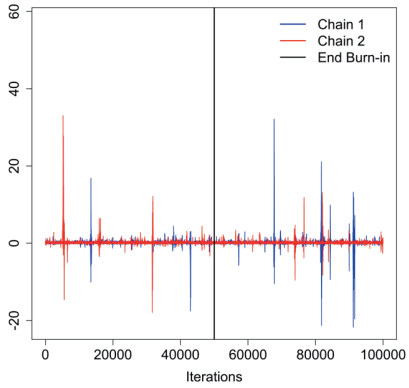
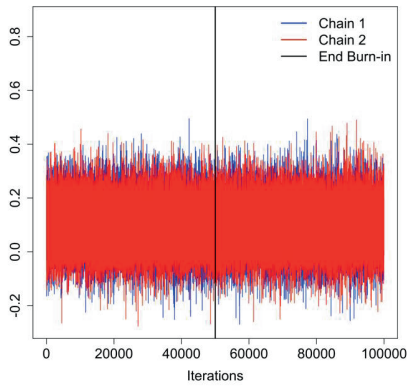
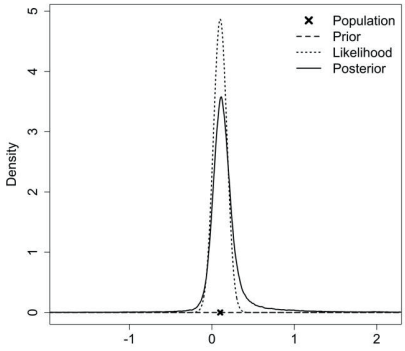
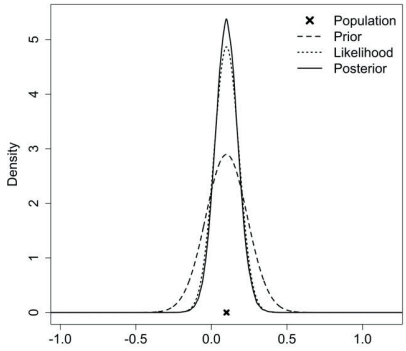
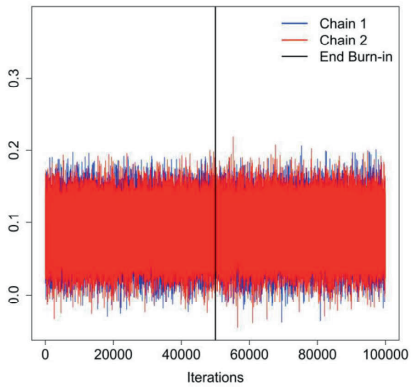
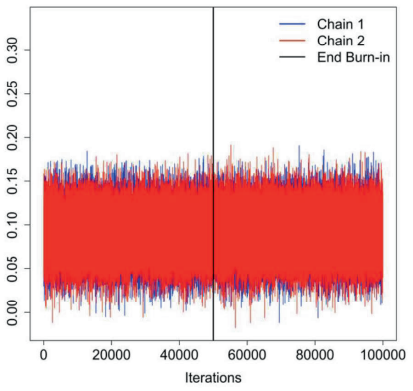
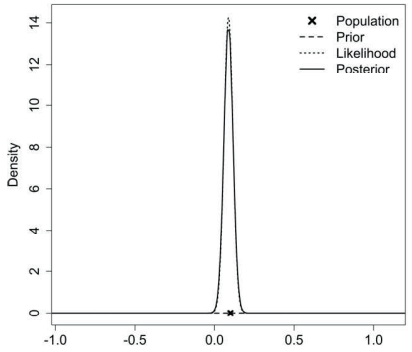
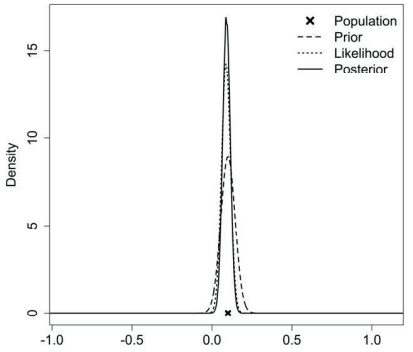


Figure 4.5: Trace plots; prior, likelihood, posterior plots; posterior standard deviation (SD) and 95% highest posterior density intervals (HPD) for regression coefficient $\beta 2$ under sample size n = 26 when M*plus* default priors and partial thoughtful priors are specified. Figures under sample sizes n = 52 and 325 are shown on the next pages.

N = 52

M*plus* Default priors	Partial Thoughtful Priors
E 	**F** 
G 	**H** 
Posterior SD = 1.01 95% HPD = -0.27; 0.93	Posterior SD = 0.08 95% HPD = -0.05; 0.25

N = 325	
M*plus* Default priors	Partial Thoughtful Priors
I 	J 
K 	L 
Posterior SD = 0.03 95% HPD = 0.03; 0.14	Posterior SD = 0.02 95% HPD = 0.04; 0.14

4.7 Summary

In this tutorial, we discussed the risks associated with default priors in Bayesian SEM when samples are small. We described the *dangers of the defaults* as a combination of three factors: (a) the relatively larger impact of the prior on the posterior when samples are small, (b) the wide distribution of default priors that often contain unrealistic values, and (c) the *false belief* that default priors are noninformative priors. We demonstrated an interactive Shiny app, in which users can investigate the impact of priors and sample size on model results. The Shiny app can also be used to teach students about responsible use of Bayesian SEM with small samples. In this chapter, we showed that default priors can *act* as highly informative priors when samples are small. We provided an overview of relevant literature (available via https://osf.io/pnmde) on how to specify suitable priors based on multiple sources of information. We discussed how to recognize a misbehaving prior by inspecting (a) effective sample size, (b) trace plots, (c) the comparison of prior-likelihood-posterior distributions, (d) posterior standard deviation and 95% highest posterior densities.

It is important to note that we are not arguing that researchers are solely responsible for breaking away from their reliance on default priors. There are several strategies that could be employed to help researchers improve their decisions regarding prior specification. A simple way in which the use of Bayesian methods can be improved is by making available educational tools, such as the App introduced in this chapter, to a broad audience of researchers. More generally, software developers could implement notifications that nudge users to check the impact of their prior distributions through techniques proposed in the current chapter (e.g., flag low ESSs and suggest inspection of trace plots). Another opportunity to intervene and improve occurs during the peer-review process. Reviewers should closely examine the decisions authors have made regarding their prior specification and intervene if the decisions made by the authors were inappropriate. In such a case, a reviewer can advise that revisions are in order to ensure that Bayesian methods were applied appropriately.

Bayesian SEM should only be used with small samples when information is available about the reasonable range of values for all parameters in the model. This information is necessary to formulate suitable prior distributions *and* to check afterwards whether the priors are 'behaving'. It is our hope that this tutorial helps spread awareness that the use of Bayesian estimation is not a *quick solution* to small sample problems in SEM, and that we encourage researchers to specify suitable prior distributions and carefully check the results when using Bayesian SEM with small samples.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, *103*(3), 411–423.

Ashby, D. (2006). Bayesian statistics in medicine: A 25 year review. *Statistics in Medicine*, *25*(21), 3589–3631. doi:10.1002/sim.2672

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

Asparouhov, T., & Muthén, B. O. (2014). Auxiliary Variables in Mixture Modeling: Three-Step Approaches Using M*plus*. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329–341. doi:10.1080/10705511.2014.915181

Bainter, S. A. (2017). Bayesian Estimation for Item Factor Analysis Models with Sparse Categorical Indicators. *Multivariate Behavioral Research*, *52*(5), 593–615. doi:10.1080/00273171.2017.1342203

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2016). Relating latent class membership to continuous distal outcomes: Improving the LTB approach and a modified three-step implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 278–289.

Bakk, Z., Oberski, D., & Vermunt, J. (2014). Relating latent class assignments to external variables: Standard errors for corrected inference. *Political Analysis*, *22*, 520–540.

Bakk, Z., & Vermunt, J. K. (2016). Robustness of Stepwise Latent Class Modeling With Continuous Distal Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 20–31. doi:10.1080/10705511.2014.955104

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*(2), 151–164. doi:10.1037/a0030642

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. doi:10.1016/j.paid.2006.09.018

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. doi:10.1037/1082-989X.8.3.338

Berger, J. O., & Bayarri, M. J. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, *19*(1), 58–80. doi:10.1214/088342304000000116

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, *37*(2), 905–938. doi:10.1214/07-AOS587

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 113–147.

Betancourt, M. (2017). How the Shape of a Weakly Informative Prior Affects Inferences. Retrieved from https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html

Bolsinova, M., Hoijtink, H., Vermeulen, J. A., & Béguin, A. (2017). Using expert knowledge for test linking. *Psychological Methods*, *22*(4), 705–724. doi:10.1037/met0000124

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., ... Hughes, C. (2014). Reaching the hard-to-reach: A systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology*, *14*(1), 1–29.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. doi:10.1007/BF02294248

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating Bias in Classify-Analyze Approaches for Latent Class Analysis. *Structural Equation Modeling : A Multidisciplinary Journal*, *22*(1), 1–11. doi:10.1080/10705511.2014.935265

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*(3), 391–420. Retrieved from https://users.soe.ucsc.edu/~draper/browne-draper-2000.pdf

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473–514.

Burt, R. S. (1976). Interpretational Confounding of Unobserved Variables in Structural Equation Models. *Sociological Methods & Research*, *5*(1), 3–52. doi:10.1177/004912417600500101

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Thomson Learning.

Chen, J., Choi, J., Weiss, B. A., & Stapleton, L. (2014). An Empirical Evaluation of Mediation Effect Analysis with Manifest and Latent Variables Using Markov Chain Monte Carlo and Alternative Estimation Methods. *Structural Equation Modeling*, *21*(2), 253–262. doi:10.1080/10705511.2014.882688

Chen, J., Zhang, D., & Choi, J. (2015). Estimation of the latent mediated effect with ordinal data using the limited-information and Bayesian full-information approaches. *Behavior Research Methods*, *47*(4), 1260–1273.

Chow, S.-M. C., & Hoijtink, H. (Eds.). (2017). Bayesian Data Analysis - Part II [Special issue]. *Psychological Methods*, *22*(4).

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.

Coleman, M. J., Cook, S., Matthysse, S., Barnard, J., Lo, Y., Levy, D. L., ... Holzman, P. S. (2002). Spatial and object working memory impairments in schizophrenia patients: A Bayesian item-response theory analysis. *Journal of Abnormal Psychology*, *111*(3), 425–435. doi:10.1037//0021-843X.111.3.425

Croon, M. (2002). Using Predicted Latent Scores in General Latent Structure Models. In G. Marcoulides & I. Moustaki (Eds.), *Latent Variable and Latent Structure Models* (pp. 207–236). Mahwah, NJ: Lawrence Erlbaum. doi:10.4324/9781410602961-16

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve Frequently Asked Questions About Growth Curve Modeling. *Journal of Cognition and Development : Official Journal of the Cognitive Development Society*, *11*(2), 121–136. doi:10.1080/15248371003699969

Darnieder, W. F. (2011). Bayesian methods for data-dependent priors (Doctoral dissertation).

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., ... Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review of Educational Research*, *79*(1), 69–102. doi:10.3102/0034654308325581

Depaoli, S. (2012). Measurement and Structural Model Class Separation in Mixture CFA: ML/EM Versus MCMC. *Structural Equation Modeling*, *19*(2), 178–203. doi:10.1080/10705511.2012.659614

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

Depaoli, S. (2014). The impact of "inaccurate" informative priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, *21*, 239–252.

Depaoli, S. (2021). *Bayesian Structural Equation Modeling*. The Guilford Press.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes. *Structural Equation Modeling*, *22*(3), 327–351. doi:http://dx.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. doi:https://dx.doi.org/10.1037/met0000065

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis Testing Using Factor Score Regression: A Comparison of Four Methods. *Educational and Psychological Measurement*, *76*(5), 741–770. doi:10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor Score Path Analysis: An Alternative for SEM? *Methodology*, *13*(Supplement 1), 31–38. doi:10.1027/1614-2241/a000130

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An Introduction to*

Latent Variable Growth Curve Modeling : Concepts, Issues, and Application, Second Edition. Routledge. doi:10.4324/9780203879962

Eastman, A. L., Mitchell, M. N., & Putnam-Hornstein, E. (2016). Risk of re-report: A latent class analysis of infants reported for maltreatment. *Child Abuse & Neglect*, *55*, 22–31. doi:10.1016/j.chiabu.2016.03.002

Egberts, M. R., Van de Schoot, R., Boekelaar, A., Hendrickx, H., Geenen, R., & Van Loey, N. E. E. (2016). Child and adolescent internalizing and externalizing problems 12 months postburn: The potential role of preburn functioning, parental posttraumatic stress, and informant bias. *European Child & Adolescent Psychiatry*, *25*(7), 791–803. doi:10.1007/s00787-015-0788-z

Erosheva, E. A., & Curtis, S. M. (2011). Dealing with rotational invariance in Bayesian confirmatory factor analysis. *Technical Report. Department of Statistics, University of Washington, Seattle, Washington, USA.*, 35.

Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*(6), 1209–1217. doi:10.3758/PBR.15.6.1209

Ferdinands, G. (2021). AI-assisted systematic reviewing: Selecting studies to compare bayesian versus frequentist SEM for small sample sizes. *Multivariate Behavioral Research*, *56*(1), 153–154.

Flegal, J. M., Hughes, J., Vats, D., Dai, N., Gupta, K., & Maji, U. (2021). *Mcmcse: Monte carlo standard errors for MCMC*. Riverside, CA,; Kanpur, India.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis, Third Edition*. Boca Raton, FL, USA: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, 473–483.

Hallquist, M., & Wiley, J. (2017). MplusAutomation: Automating Mplus Model Estimation and Interpretation. Retrieved from https://CRAN.R-project.org/package=MplusAutomation

Hertzog, C., Oertzen, T. von, Ghisletta, P., & Lindenberger, U. (2008). Evaluating the Power of Latent Growth Curve Models to Detect Individual Differences in Change. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 541–563. doi:10.1080/10705510802338983

Hipwell, A. E., Stepp, S. D., Moses-Kolko, E. L., Xiong, S., Paul, E., Merrick, N., ... Keenan, K. (2016). Predicting adolescent postpartum caregiving from trajectories of depression and anxiety prior to childbirth: A 5-year prospective study. *Archives of Women's Mental Health*, *19*(5), 871–882. doi:10.1007/s00737-016-0627-3

Hoijtink, H., & Chow, S.-M. C. (Eds.). (2017). Bayesian Data Analysis - Part I [Special issue]. *Psychological Methods*, *22*(2).

Holgersen, K. H., Boe, H. J., Klöckner, C. A., Weisæth, L., & Holen, A. (2010). Initial stress responses in relation to outcome after three decades. *The Journal of Nervous and Mental Disease*, *198*(3), 230–233. doi:https://doi.org/10.1097/NMD.0b013e3181d106a9

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. doi:10.1080/00273171.2016.1208074

Hoogland, J. J., & Boomsma, A. (1998). Robustness Studies in Covariance Structure Modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367. doi:10.1177/0049124198026003003

Hoshino, T., & Bentler, P. (2013). Bias in factor score regression and a simple solution. In A. de Leon & K. Chough (Eds.), *Analysis of Mixed Data* (pp. 43–61). Chapman; Hall/CRC. doi:10.1201/b14571-5

Hox, J. J., & Maas, C. J. M. (2001). The Accuracy of Multilevel Structural Equation Modeling With Pseudobalanced Groups and Small Samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 157–174. doi:10.1207/S15328007SEM0802_1

Hox, J. J., Moerbeek, M., Kluytmans, A., & Van de Schoot, R. (2014). Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and

power. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00078

Hox, J. J., Van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*(2), 87–93.

Huang, D., Brecht, M.-L., Hara, M., & Hser, Y.-I. (2010). Influences of a Covariate on Growth Mixture Modeling. *Journal of Drug Issues*, *40*(1), 173–194. doi:https://doi.org/10.1177/002204261004000110

IBM Corp. (2017). IBM SPSS Statistics for Windows. Armonk, NY: IBM Corp.

JASP team. (2018). JASP. [Computer Software].

Jeffreys, H. (1945). An invariant form for the prior probability in estimation problems. Retrieved from http://rspa.royalsocietypublishing.org/

Jiang, L., Chen, S., Zhang, B., Beals, J., Mitchell, C. M., Manson, S. M., & Roubideaux, Y. (2016). Longitudinal Patterns of Stages of Change for Exercise and Lifestyle Intervention Outcomes: An Application of Latent Class Analysis with Distal Outcomes. *Prevention Science*, *17*(3), 398–409. doi:10.1007/s11121-015-0599-y

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

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. The Guilford Press.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In *T.D. Little (ed.), Oxford handbook of quantitative methods* (pp. 407–437). Oxford: Oxford University Press.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.

Ke, Z., & Wang, L. (2015). Detecting Individual Differences in Change: Methods and Comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 382–400. doi:10.1080/10705511.2014.936096

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood.

Computational Statistics & Data Analysis, *53*(7), 2583–2595. doi:10.1016/j.csda.2008.12.013

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.

König, C., & Van de Schoot, R. (2017). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, 1–24. doi:10.1080/00131911.2017.1350636

Koopman, J., Howe, M., Hollenbeck, J. R., & Sin, H.-P. (2015). Small sample mediation testing: Misplaced confidence in bootstrapped confidence intervals. *Journal of Applied Psychology*, *100*(1), 194–202. doi:10.1037/a0036635

Koukounari, A., Stringaris, A., & Maughan, B. (2017). Pathways from maternal depression to young adult offspring depression: An exploratory longitudinal mediation analysis. *International Journal of Methods in Psychiatric Research*, *26*(2), e1520. doi:10.1002/mpr.1520

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676. doi:10.1002/wcs.72

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). London, UK: Academic Press.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752. doi:10.1177/1094428112457829

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent Class Analysis with Distal Outcomes: A Flexible Model-Based Approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 1–26. doi:10.1080/10705511.2013.742377

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, UK: Cambridge University Press.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, West Sussex, England: John Wiley & Sons.

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*(4), 653–686. doi:https://doi.org/10.1207/s15327906mbr3904_4

Lee, S.-Y., Song, X.-Y., & Tang, N.-S. (2007). Bayesian Methods for Analyzing Structural Equation Models With Covariates, Interaction, and

Quadratic Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 404–434. doi:10.1080/10705510701301511

LeGower, M., & Walsh, R. (2017). Promise scholarship programs as place-making policy: Evidence from school enrollment and housing prices. *Journal of Urban Economics*, *101*, 74–89. doi:10.1016/j.jue.2017.06.001

Lek, K., & Van de Schoot, R. (2018). Development and evaluation of a digital expert elicitation method aimed at fostering elementary school teachers' diagnostic competence. In *Frontiers in education* (Vol. 3, p. 82). Frontiers.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York, NY, US: Guilford Press.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of Inverse Wishart and Separation-Strategy Priors for Bayesian Estimation of Covariance Parameter Matrix in Growth Curve Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 354–367. doi:10.1080/10705511.2015.1057285

Loken, E. (2005). Identification Constraints and Inference in Factor Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(2), 232–244. doi:10.1207/s15328007sem1202_3

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY, US: Springer Science & Business Media.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. Routledge.

Masyn, K., Petras, H., & Lu, W. (2014). Growth Curve Models with Categorical Outcomes. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of Criminology and Criminal Justice* (pp. 2013–2025). New York, NY: Springer New York. doi:10.1007/978-1-4614-5690-2

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics*, *16*(1), 163–200. doi:10.1007/BF01065485

McArdle, J. J. (2012). Latent curve modeling of longitudinal growth data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling.* (pp. 547–571). The Guilford Press, New York.

McArdle, J. J., & Epstein, D. (1987). Latent Growth Curves within Developmental Structural Equation Models. *Child Development*, *58*(1), 110–133. doi:10.2307/1130295

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological reserach. In W. F. Velicer & J. Schinka (Eds.), *Handbook of psychology: Research methods in psychology* (pp. 447–480). New York: Wiley.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Bioca Raton, FL: CRC Press, Taylor & Francis Group.

McNeish, D. (2016a). On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. doi:10.1080/10705511.2016.1186549

McNeish, D. (2016b). 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. doi:10.3102/1076998615621299

McNeish, D. (2017). Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research*, *52*(5), 661–670. doi:10.1080/00273171.2017.1344538

McNeish, D. (2018). Brief research report: Growth models with small samples and missing data. *The Journal of Experimental Education*, *86*(4), 690–701.

McNeish, D., & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. doi:10.1080/00273171.2016.1167008

Meng, X.-L., & Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, *80*(2), 267–278. doi:10.2307/2337198

Meredith, W., & Tisak, J. (1990). Latent Curve Analysis. *Psychometrika*, *55*(1), 107–122.

Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of Statistical Software*, *85*(4), 1–30.

Miočević, M., MacKinnon, D. P., & Levy, R. (2017). Power in Bayesian Mediation Analysis for Small Sample Research. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 666–683. doi:10.1080/10705511.2017.1312407

Mitani, A. A., & Haneuse, S. (2020). Small Data Challenges of Studying

Rare Diseases. *JAMA Network Open*, *3*(3), e201965–e201965. doi:10.1001/jamanetworkopen.2020.1965

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. doi:10.1371/journal.pmed.1000097

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

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide.* (Eight edition.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. doi:10.1207/S15328007SEM0904_8

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. doi:10.1016/S0022-2496(02)00028-7

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.01599

Nevitt, J., & Hancock, G. R. (2004). Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modeling. *Multivariate Behavioral Research*, *39*(3), 439–478. doi:10.1207/S15327906MBR3903_3

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities.* John Wiley & Sons.

Petras, H., & Masyn, K. (2010). General Growth Mixture Analysis with Antecedents and Consequences of Change. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 69–100). New York, NY: Springer New York. doi:10.1007/978-0-387-77650-7_5

Polson, N. G., & Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*, 887–902.

Price, L. R. (2012). Small sample properties of bayesian multivariate autoregressive time series models. *Structural Equation Modeling*, *19*(1), 51–64.

R Core Team. (2022). *R: A language and environment for statistical*

computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rietbergen, C., Debray, T. P. A., Klugkist, I., Janssen, K. J. M., & Moons, K. G. M. (2017). Reporting of Bayesian analysis in epidemiologic research should become more transparent. *Journal of Clinical Epidemiology*, *86*, 51–58.e2. doi:10.1016/j.jclinepi.2017.04.008

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Rosseel, Y. (2020). Small sample solutions for structural equation modelling. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge.

Rosseel, Y., & Devlieger, I. (2018, March). Why we may not need SEM after all. Meeting of the SEM Working Group, Amsterdam.

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement (SAM) approach to structural equation modeling. Retrieved from https://osf.io/pekbm/

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), 424–451.

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models. *Multivariate Behavioral Research*, *51*(2-3), 185–206. doi:10.1080/00273171.2015.1065398

Serang, S., Zhang, Z., Helm, J., Steele, J. S., & Grimm, K. J. (2015). Evaluation of a Bayesian Approach to Estimating Nonlinear Mixed-Effects Mixture Models. *Structural Equation Modeling*, *22*(2), 202–215. doi:10.1080/10705511.2014.937322

Shi, D., & Tong, X. (2017). The Impact of Prior Information on Bayesian Latent Basis Growth Model Estimation. *SAGE Open*, *7*(3), 215824401772703. doi:10.1177/2158244017727039

Shin, T., Davison, M. L., & Long, J. D. (2017). Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychological Methods*, *22*(3), 426–449. doi:10.1037/met0000094

Smid, S. C., Depaoli, S., & Van de Schoot, R. (2020). Predicting a distal outcome variable from a latent growth model: ML versus bayesian

estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(2), 169–191.

Smid, S. C., McNeish, D., Miočević, M., & Van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161.

Smid, S. C., & Winter, S. D. (2020). Shiny App: The Impact of Prior Distributions in a Bayesian Latent Growth Model. Retrieved from https://osf.io/m6byv/

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, *4*(38).

Spoth, R., Clair, S., & Trudeau, L. (2014). Universal Family-Focused Intervention with Young Adolescents: Effects on Health-Risking Sexual Behaviors and STDs Among Young Adults. *Prevention Science*, *15*(1), 47–58. doi:10.1007/s11121-012-0321-2

Stan Development Team. (2017). Stan Modeling Language: User's Guide and Reference Manual. Version 2.17.0.

Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science*, *57*(3), 748–761. doi:10.1111/ajps.12001

Stoel, R. D., Wittenboer, G., van den, & Hox, J. J. (2004). Methodological Issues in the Application of the Latent Growth Curve Model. In.

Takane, Y., & Hwang, H. (2018). Comparisons among several consistent estimators of structural equation models. *Behaviormetrika*. doi:10.1007/s41237-017-0045-5

Tong, X., & Ke, Z. (2016). Growth Curve Modeling for Nonnormal Data: A Two-Stage Robust Approach Versus a Semiparametric Bayesian Approach. *Quantitative Psychology Research*, *167*, 229–241. doi:10.1007/978-3-319-38759-8_17

Tsai, M.-Y., & Hsiao, C. K. (2008). Computation of reference Bayesian inference for variance components in longitudinal studies. *Computational Statistics*, *23*(4), 587–604. doi:10.1007/s00180-007-0100-x

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1), 1–10.

Van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & Loey, N. E. van. (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). doi:10.3402/ejpt.v6.25216

Van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist*, *16*(2), 75–84.

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

Van de Schoot, R., & Miočević, M. (2020). *Small sample size solutions: A guide for applied researchers and practitioners*. Taylor & Francis.

Van de Schoot, R., Schalken, N., & Olff, M. (2017). Systematic search of Bayesian statistics in the field of psychotraumatology. *European Journal of Psychotraumatology*, *8*(sup1). doi:10.1080/20008198.2017.1375339

Van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., & Loey, N. E. van. (2018). Bayesian PTSD-Trajectory Analysis with Informed Priors Based on a Systematic Literature Search and Expert Elicitation. *Multivariate Behavioral Research*, *53*(2), 267–291. doi:10.1080/00273171.2017.1412293

Van de Schoot, R., Sijbrandij, M., Winter, S. D., Depaoli, S., & Vermunt, J. K. (2016). The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–17. doi:10.1080/10705511.2016.1247646

Van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A Systematic Review of Bayesian Papers in Psychology: The Last 25 Years, *22*(2), 217–239. doi:http://dx.doi.org/10.1037/met0000100

Van Erp, S. J., Mulder, J., & Oberski, D. L. (2018). Prior Sensitivity Analysis in Default Bayesian Structural Equation Modeling. *Psychological Methods*, *23*(2), 363–388.

Van Lier, H. G., Oberhagemann, M., Stroes, J. D., Enewoldsen, N. M., Pieterse, M. E., Schraagen, J. M. C., ... Noordzij, M. L. (2017). Design Decisions for a Real Time, Alcohol Craving Study Using Physio- and Psychological Measures. In *De Vries, P.W., Oinas-Kukkonen, H.,*

Siemons, L., Beerlage-de Jong, N., & van Gemert-Pijnen, L. (Eds.). Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors. 12th International Conference, PERSUASIVE 2017, Amsterdam, The Netherlands, April 4–6, 2017, Proceedings.

Vats, D., Flegal, J. M., & Jones, G. L. (2019). Multivariate output analysis for markov chain monte carlo. *Biometrika*, *106*(2), 321–337.

Veen, D., & Egberts, M. (2020). The importance of collaboration in bayesian analysis with small samples. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners.* Routledge.

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*, 1197.

Veen, D., Stoel, D., Zondervan-Zwijnenburg, M., & Van de Schoot, R. (2017). Proposal for a Five-Step Method to Elicit Expert Judgment. *Frontiers in Psychology*, *8*. doi:10.3389/fpsyg.2017.02110

Vermunt, J. K. (2010). Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Political Analysis*, *18*(4), 450–469. doi:https://doi.org/10.1093/pan/mpq025

Vollebregt, S. J., Scholte, W. F., Hoogerbrugge, A., Bolhuis, K., & Vermeulen, J. M. (2022). Help-seeking undocumented migrants in the netherlands: Mental health, adverse life events, and living conditions. *Culture, Medicine, and Psychiatry*, 1–23.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY, US: Springer Science + Business Media. doi:10.1007/978-0-387-09612-4_9

Wickham, H. (2016). Package "ggplot2": Elegant graphics for data analysis. *Springer-Verlag New York*.

Wijnen, F., Bree, E. de, Alphen, P. van, Jong, J. de, & Leij, A. van der. (2015). Comparing SLI and dyslexia: Developmental language profiles and reading outcomes. In S. Stavrakaki (Ed.), *Specific Language Impairment: Current trends in research* (pp. 89–112). John Benjamins Publishing Company.

Yang, R., & Berger, J. O. (1996). *A catalog of noninformative priors.* Institute

of Statistics; Decision Sciences, Duke University.

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge Structural Equation Modeling with Correlation Matrices for Ordinal and Continuous Data. *The British Journal of Mathematical and Statistical Psychology*, *64*(0 1). doi:10.1348/000711010X497442

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301–322. doi:10.1037/a0016972

Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*(4), 374–383.

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*, *0*(0), 1–16. doi:10.1080/10705511.2018.1545232

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the Performance of Bayesian Approaches in Small Samples: A comment on Smid, McNeish, Miočević, and Van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 40–50.

Zondervan-Zwijnenburg, M., Depaoli, S., Peeters, M., & Van de Schoot, R. (2019). Pushing the Limits: The Performance of Maximum Likelihood and Bayesian Estimation With Small and Unbalanced Samples in a Latent Growth Model. *Methodology*, *1*(1), 1–13. doi:10.1027/1614-2241/a000162

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*, *14*(4), 305–320. doi:10.1080/15427609.2017.1370966

Nederlandse Samenvatting

Tegenwoordig lijkt het alsof er over alles meer dan genoeg data beschikbaar is. Echter, zelfs in tijden van *big data* zijn er situaties waarin het uitdagend is om genoeg data te verzamelen. Denk aan kleine populaties, zoals mensen met zeer zeldzame ziektes. Of een groep die moeilijk te bereiken is, zoals mensen met verslavingsproblemen, laaggeletterden, of ongedocumenteerde migranten. Ook financiële redenen kunnen een rol spelen, denk aan studies waarin dure MRI scans gebruikt worden en genoeg data verzamelen simpelweg te veel kost. Kleine steekproeven zijn in dit soort situaties onvermijdelijk.

Structurele vergelijkingsmodellen (*Structural Equation Models*; SEMs) hebben net als alle statistische modellen, een bepaalde hoeveelheid data nodig om goed te functioneren. Kleine steekproeven kunnen grote problemen veroorzaken. Zoals een model dat niet convergeert, afwijkende of nietszeggende resultaten genereert. Eén manier om deze problemen te vermijden is het versimpelen van de onderzoeksvraag en het statistische model. Dit is echter niet wenselijk, omdat complexe en essentiële onderzoeksvragen over moeilijk te bereiken groepen en kleine populaties dan niet beantwoord kunnen worden, en er zo belangrijke informatie misgelopen wordt.

Een andere manier om de kleine-steekproef-problemen te omzeilen is door het gebruik van Bayesiaanse statistiek. In theorie is het Bayesiaanse framework meer geschikt voor kleine datasets, in vergelijking met de klassieke frequentistische methoden (zoals *Maximum Likelihood* schatting). Allereerst doordat met het specificeren van prior verdelingen extra informatie toegevoegd kan worden aan de analyse. Daarnaast zijn Bayesiaanse methoden niet gebaseerd op *large-sample techniques*.

In de praktijk zien we dat onderzoekers steeds vaker kiezen voor het gebruik van Bayesiaanse SEM (BSEM), om zo kleine-steekproef-problemen te omzeilen. De overstap naar Bayesiaanse statistiek is echter niet zonder problemen. Wanneer BSEM gebruikt wordt, moeten er prior verdelingen gespecificeerd worden voor alle parameters in het model. Hoe kleiner de steekproef, hoe groter de impact van de prior verdelingen op de posterior. Er zijn een aantal software programma's die gebruik maken van ingebouwde *default prior* verdelingen. Helaas zijn deze priors niet altijd geschikt en kunnen ze tot onjuiste resultaten leiden in het geval van kleine steekproeven.

Het doel van dit proefschrift is dan ook allereerst het creëren van overzicht over het functioneren van Bayesiaanse schattingsmethoden voor SEM met kleine steekproeven. Daarnaast bespreken we voorzorgsmaatregelen en verstrekken we richtlijnen zodat BSEM op een *bewuste* manier gebruikt kan worden met kleine steekproeven.

Elk hoofdstuk heeft een pagina op het Open Science Framework (OSF). Hier zijn alle aanvullende bestanden te vinden, inclusief geannoteerde R en M*plus* code om te resultaten te reproduceren. De link naar de bijbehorende OSF pagina wordt gegeven aan het begin van elk hoofdstuk.

In **hoofdstuk 1**, bespreken we de resultaten van een uitgebreide literatuurstudie. We geven een overzicht van 32 simulatiestudies, waarin de werking van Bayesiaanse en frequentistische methoden wordt onderzocht voor SEMs met kleine steekproeven. Een reeks aan verschillende SEMs wordt besproken, wat de resultaten van deze studie breed toepasbaar maakt. We presenteren een overzicht van de geïncludeerde studies, laten zien wat volgens de auteurs van de studies beschouwd wordt als een kleine steekproef, en voegen de informatie uit de studies samen in figuren. Op basis van de literatuurstudie, concluderen we dat met kleine steekproeven het gebruik van Bayesiaanse schattingsmethoden met *default priors* kan leiden tot ernstig afwijkende resultaten. We eindigen het hoofdstuk met aanbevelingen om bewuste keuzes te maken over alle prior verdelingen in het model, en geven voorbeelden hoe deze prior verdelingen opgesteld kunnen worden.

In **hoofdstuk 2** onderzoeken we het latente groei model (LGM) met een lange-termijn uitkomstmaat in een simulatiestudie. Bayesiaanse schatting (met M*plus default* en informatieve priors) wordt vergeleken met *Maximum Likelihood* (ML) schatting. Voorzichtigheid is geboden wanneer (1) de

steekproef klein is; en (2) er weinig variantie rondom de latente slope wordt verwacht, zelfs wanneer de steekproef groot is. We adviseren om ML en Bayesiaanse schatting met M*plus* *default priors* niet te gebruiken in bovenstaande situaties om ernstig afwijkende resultaten te voorkomen. De specificatie van informatieve priors kan uitkomst bieden. Aan de andere kant, de specificatie van priors die afwijken van de populatiewaarden, verslechteren de resultaten bij een kleine steekproef. Wij adviseren onderzoekers daarom om de meest zorgvuldige benadering te gebruiken: begin met het zorgvuldig specificeren van prior verdelingen; en onderzoek de impact en robuustheid van de priors naderhand in een uitgebreide sensitiviteitsanalyse.

In **hoofdstuk 3** bespreken we twee veelbelovende frequentistische methoden om SEM met kleine steekproeven te analyseren: *twostep modeling* (twostep) en *Factor Score Regression* (FSR). In het hoofdstuk worden deze twee methoden uitgelegd, en in een simulatiestudie vergeleken met *Maximum Likelihood* (ML) schatting, Bayesiaanse schatting met `blavaan` *default priors* en informatieve priors voor verschillende steekproefgroottes. Met kleine steekproeven, geven bijna alle methoden problemen. De frequentistische methoden (ML, twostep, FSR) in termen van een model dat niet convergeert, negatieve varianties en extreme parameter schattingen; en de Bayesiaanse methode met *default priors* in termen van *mode-switching* en *spikes*. Wanneer het vergroten van de steekproef geen optie is, raden wij aan Bayesiaanse statistiek te gebruiken met informatieve priors. Wanneer onderzoekers geen prior informatie willen of kunnen toevoegen, adviseren wij twostep of FSR te gebruiken. Deze methoden zijn een veiligere keuze dan ML: ze leiden vaker tot een convergerend model zonder negatieve varianties, stabielere resultaten tussen replicaties in de simulatiestudie en minder extreem afwijkende parameter schattingen dan ML met kleine steekproeven.

De eerste drie studies in dit proefschrift duiden op dezelfde conclusie: overstappen naar een Bayesiaanse schattingsmethode en daarbij blind vertrouwen op de ingebouwde *default priors* is geen oplossing. Met kleine steekproeven kan het gebruik van *default priors* leiden tot onjuiste parameter schattingen (vooral ernstig afwijkende variantie parameter schattingen), instabiele resultaten, en grote onzekerheid in de posterior verdeling.

In **hoofdstuk 4**, brengen we op een toegankelijke manier de risico's van het gebruik van Bayesiaanse statistiek met *default priors* en kleine steekproeven onder de aandacht. We bespreken de relatief grote impact van de prior op de posterior wanneer steekproeven klein zijn, het probleem met de vaak zeer wijde verdeling van de ingebouwde *default priors*, en de onjuiste overtuiging dat *default priors* niet-informatieve priors zouden zijn en geen impact zouden hebben op de resultaten. Daarnaast presenteren we een online shiny applicatie (Smid & Winter, 2020, beschikbaar via https://osf.io/m6byv/), waarbij gebruikers de impact van verschillende priors en steekproefgroottes op de resultaten kunnen exploreren. We bespreken hoe de app gebruikt kan worden in het onderwijs, en we verstrekken een lijst met literatuur over het specificeren van prior verdelingen. We sluiten af met richtlijnen over het herkennen van *misbehaving* en *behaving* priors na het uitvoeren van een Bayesiaanse analyse.

Dankwoord

Daar is die dan, mijn proefschrift! Ik ben trots, en ik kan bijna niet geloven dat het nu echt allemaal klaar is. Graag wil ik een aantal mensen bedanken die direct en indirect hebben bijgedragen aan het onderzoek in dit proefschrift.

Ik begin met mijn promotoren. Rens, bedankt voor de vrijheid, het vertrouwen en de verantwoordelijkheid die ik heb gekregen de afgelopen jaren. Ik wil je bedanken voor je enthousiasme, je enorme hoeveelheid energie, en alle mogelijkheden en kansen tijdens mijn PhD project. Je gaf me de ruimte om mezelf te ontwikkelen. Dankjewel daarvoor.

Leoniek, ik wil je bedanken voor je steun, luisterend oor, en de waardevolle gesprekken die we de afgelopen jaren hebben gehad, en nog steeds hebben. Het betekende veel dat ik bij jou terecht kon toen het met mij minder goed ging. Bedankt voor ons fijne contact.

Of course I would also like to thank my co-authors Dan McNeish, Milica Miočeviç, Sarah Depaoli, Yves Rosseel and Sonja Winter. Thank you for the fun collaboration, inspiration and contributions to the research in this thesis. Sarah, thank you for hosting me twice at the University of California in Merced. Also a special thanks to Sonja and Amber for showing me around and making me feel so welcome at UC Merced.

Ook wil ik student-assistenten Naomi Schalken, Gerbrich Ferdinands en Laura Hofstee bedanken. Hartelijk bedankt voor jullie bijdrage aan dit proefschrift.

Aan alle oud-M&S-collega's van de Universiteit Utrecht: vanaf 2012 liep ik op de afdeling rond als masterstudent, later als junior docent, en daarna als promovendus. Wat een fijne werkomgeving is dit! Ik wil iedereen bedanken voor de fijne sfeer, koffies, de meetings met inhoudelijke discussies, en de borrels zonder inhoudelijke discussies. Dank aan Kees, Frank, Corine, Fayette, Duco, Anne, Hidde, Kimberley, Mariëlle, Karlijn, en alle andere fijne, en gezellige collega's!

Joop, ik werd student-assistent bij jou tijdens de onderzoeksmaster. Mijn eerst gepubliceerde paper ooit is het paper waar wij samen aan gewerkt hebben. Bedankt voor de leuke samenwerking, en alles wat ik van je geleerd heb.

Kevin, bedankt voor alle hulp bij het regelen van allerlei praktische dingen (inclusief het redden van een cursus jaren geleden). Marianne, dank voor de gezelligheid tijdens conferenties.

Een speciaal bedankje aan Corine en Duco: wat een geluk om in de 'gezelligste kamer van de UU' te zitten tijdens mijn PhD (inclusief kameruitjes en geheime tosti's).

Ook iedereen van de M&S master: bedankt voor de gezelligheid tijdens het samen studeren, het zoveel leren van elkaar, en het leuke contact.

Aan alle oud-collega's van bakkerij Loaf, lieve loafies, bedankt voor alle warmte, de fijne werkplek, de overheerlijke zuurdesembroden en vegan croissants, en alles wat ik heb geleerd in mijn tijd als bakker!

Lieve Eva en Bente, bedankt dat jullie mijn paranimfen zijn! Maar nog veel meer bedankt voor de fijne vriendschap.

Lieve Anne, Bente, Eva, Rinske, Carlijn, Aart, Kees, Lisa, Corine, Frank, Fayette, Marit, Femke en Fionna, en alle andere lieve mensen, vrienden en familie: Ik ben zo dankbaar voor alle liefde en steun die ik de afgelopen jaren heb mogen ontvangen. Bedankt voor alle gezelligheid, fijne gesprekken, wandelingen, spelletjes, lieve emails, etentjes, videogesprekken en kaartjes.

Lieve mama, papa, Rinske en Carlijn, bedankt voor jullie warmte, liefde, betrokkenheid en interesse. Ik voel me zo dankbaar voor jullie steun, het meeleven, het luisteren, en het vertrouwen. Bedankt dat jullie er altijd voor me zijn!

Lieve Herre, bedankt voor al die keren dat je tegen me hebt gezegd dat het wel goed zou komen, bedankt voor je geduld, alle kattenfilmpjes, en dat alles er altijd mag zijn.

En als laatste, maar zeker niet het minst belangrijk, dank aan Bobbie, Krekel en Patsy! Zonder katten zou het leven een stuk minder leuk zijn.

Sanne Smid, maart 2023

About the Author

Sanne is 33 and lives in Utrecht. She loves to spend time in nature, and she likes to read, do yoga and play board games.

Sanne was born on November 7, 1989, in Utrecht. She started in 2009 as a student at Utrecht University. In 2012, she finished the bachelor Interdisciplinary Social Sciences and in 2014 she graduated from the Research Master Methodology and Statistics of Behavioral and Social Sciences. During the research master, Sanne discovered her love for statistics, in particular Structural Equation Modeling, and teaching statistics. Consequently, she worked as a teacher at the Methodology & Statistics department at Utrecht University for 1.5 year after graduation. She taught various courses from bachelor to postgraduate level, and assisted in several research projects. In January 2016, she started her PhD project on Bayesian Structural Equation Models with small samples. During her PhD project, she visited the research group of prof. dr. Sarah Depaoli at UC Merced for a two- and one-month visit respectively, and worked on the projects described in Chapters 2 and 4 of this dissertation. In 2018, she won the best presentation award at the Small Sample Size Conference in Utrecht.

In 2021, Sanne became a baker at the sourdough bakery Loaf in Utrecht. She loved to learn everything about baking sourdough bread. In November 2022, she left the bakery to finish her dissertation. She is still baking bread and vegan cakes at home.

As of February 2023, Sanne works as a data scientist at Dienst Uitvoering Onderwijs (DUO) in The Hague.

Publications

Smid, S. C., McNeish, D., Miočeviç, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 131-161.

Smid, S. C., Depaoli, S., & Van De Schoot, R. (2020). Predicting a distal outcome variable from a latent growth model: ML versus Bayesian estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(2), 169-191.

Smid, S. C., & Rosseel, Y. (2020). SEM with small samples: Twostep modeling and factor score regression versus Bayesian estimation with informative priors. In R. Van de Schoot & M. Miočeviç (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*: Routledge.

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*.

Smid, S. C., & Winter, S. D. (2020). Shiny App: The Impact of Prior Distributions in a Bayesian Latent Growth Model. Available via https://osf.io/m6byv/

Smid, S. C., Hox, J. J., Heiervang, E. R., Stormark, K. M., Hysing, M., & Bøe, T. (2020). Measurement equivalence and convergent validity of a mental health rating scale. *Assessment, 27*(8), 1901-1913.

Duinhof, E., **Smid, S. C.**, Vollebergh, W. A. M., & Stevens, G. W. J. M. (2020). Immigration background and adolescent mental health problems: The role of family affluence, adolescent educational level and gender. *Social psychiatry and psychiatric epidemiology, 55*(4), 435-445.

Jensen, M., **Smid, S. C.**, & Bøe, T. (2020). Characteristics of adolescent boys who have displayed harmful sexual behaviour (HSB) against children of younger or equal age. *BMC psychology, 8*(1), 1-13.

Tanniou, J., **Smid, S. C.**, van der Tweel, I., Teerenstra, S., & Roes, K. C. (2019). Level of evidence for promising subgroup findings: The case of trends and multiple subgroups. *Statistics in Medicine, 38*(14), 2561-2572.