Bayesian versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review

Sanne C. Smid[1], Daniel McNeish[2], Milica Miočević[1], and Rens van de Schoot[1,3]

[1] Utrecht University, The Netherlands

[2] Arizona State University, USA

[3] North-West University, South Africa

This is an accepted manuscript of an article by Taylor & Francis in *Structural Equation Modeling: A Multidisciplinary Journal*. The final article will be available, upon publication, via its DOI: 10.1080/10705511.2019.1577140

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.

Keywords: small samples; structural equation models; systematic review; informative priors

Supplemental files: https://osf.io/7mght/

Bayesian versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review

Author Notes

Sanne C. Smid, and Milica Miočević, Department of Methodology and Statistics, Utrecht University, The Netherlands; Daniel McNeish, Psychology Department, Arizona State University, USA; Rens van de Schoot, Department of Methodology and Statistics, Utrecht University, The Netherlands, and North-West University, Optentia Research Program, Faculty of Humanities, South Africa.

This research was supported by a grant from the Netherlands organization for scientific research: NWO-VIDI-452-14-006.

Correspondence should be addressed to Sanne Smid, Department of Methodology and Statistics, Utrecht University, P.O. box 80140, 3508 TC Utrecht, The Netherlands. E-mail: s.c.smid@uu.nl

Part of this work has been presented at the Multilevel Conference in 2017, Modern Modeling Methods conference in 2017, European Conference on Developmental Psychology in 2017, Small Sample Size Solutions conference in 2018.

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.

Acknowledgements

The authors 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.

Bayesian versus Frequentist Estimation for Structural Equation Models in Small Sample Contexts: A Systematic Review

The use of Bayesian estimation is on the rise in many scientific fields (König & van de Schoot, 2017; J. K. Kruschke, Aguinis, & Joo, 2012; Rietbergen, Debray, Klugkist, Janssen, & Moons, 2017; Rupp, Dey, & Zumbo, 2004; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 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 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 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, Carlin, & Stern, 2013, 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]

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., 2013, 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, D., 2016b; van de Schoot, Schalken, & Olff, 2017; van de Schoot, Winter, et al., 2017). Popular software programs, such as: M*plus* (L. K. Muthén & Muthén, 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, 2016b). 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 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]

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.

Organization of the Paper

The remainder of the paper 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.

Methods

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, CFA, 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.

Search Strategy

Three approaches containing six searches, were conducted to identify possibly relevant papers, as displayed in Figure 1. As the first approach, we used the simulation study papers on small samples which were identified by the systematic review of van de Schoot 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) 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 et al. (2017), which was used as the first approach, included studies published until 2015. The exact search strings can be found in Supplemental File S1. 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 2, a summary of the flow charts can be found following Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA; Liberati et al., 2009; 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.

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), CFA (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. 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 3 to 5.

Bayesian vs. Frequentist Methods in Included Studies

In the current paper, 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, 2016a). 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 Appendix Table A1, 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 paper 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.

Results

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 3 to 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. 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 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 and 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 et al. (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 74.0% 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 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 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, & van Loey, 2015; Zondervan-Zwijnenburg, Depaoli, Peeters, & Van de Schoot, 2018). 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 5, the relative bias for the structural parameters (Figure 5a) and variance parameters (Figure 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 5a), and the median of the distribution of ML estimation for the variance parameters (Figure 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 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 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 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 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.

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, 2016a; van de Schoot et al., 2015). The remaining studies show that the conclusion depends on: the choice of the naive prior distribution (McNeish, 2016b; McNeish & Stapleton, 2016; e.g., McNeish and 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, 2016b; Miočević et al., 2017; e.g., McNeish (2016b) 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; McNeish, 2016b; Holtmann et al., 2016). As stated by McNeish (2016b) "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 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 5, the variance parameters are more often problematic in terms of bias than the structural parameters. Interestingly, Tsai and 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 and Sun (2009), Bernardo (1979), Jeffreys (1945) and Yang and 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 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, 2016b; 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 Appendix Table A1), a relative bias of 84.4% 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 (see e.g., Browne & Draper, 2000; Browne & Draper, 2006; Hox et al., 2012; Stegmueller, 2013; Holtmann et al., 2016) although the highest levels of relative bias were reported for the between-level covariate parameter in the study by Depaoli and Clifton (2015). The estimates for the covariate of BayesN (referred to as "noninformative (diffuse) priors" in Appendix Table A1) 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 and 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 Appendix Table A1,

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 and 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 Appendix Table A1).

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 & Song, 2004; McNeish, 2016a; 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 and Muthén (2010a). 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.

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; Natesan, 2015; Price, 2012; Serang, Zhang, Helm, Steele, & Grimm, 2015; van de Schoot et al., 2015; Yuan & MacKinnon, 2009; 2 simulation studies in Zondervan-Zwijnenburg et al., 2018; Miočević et al., 2017). 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; Depaoli, 2013; 2 simulation studies in McNeish, 2016b; Holtmann et al., 2016), 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 (2016b) 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 Appendix Table A1) 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 Appendix Table A1) performs best, followed by ML and then followed by BayesT with "weak priors" (as referred to in Appendix Table A1). Depaoli (2013) investigated 4 types of BayesT priors (referred to as "informative accurate", "weakly informative", "partial informative" and "informative and inaccurate" priors in Appendix Table A1), 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 and Fellingham (2013) concluded that BayesT with Gamma priors for the variance parameters (referred to as "thoughtful priors" in Appendix Table A1) 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 Appendix Table A1). 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., 2 simulation studies in Depaoli, 2012; Depaoli & Clifton, 2015; Depaoli, 2013; van de Schoot et al., 2015; 2 simulation studies in Zondervan-Zwijnenburg et al., 2018; Holtmann et al., 2016), or by adjusting both hyperparameters (2 simulation studies in McNeish, 2016b; 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 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, 2016b). However, the results can further be improved by adding *stronger* prior information (e.g., Depaoli, 2012; McNeish, 2016b; Holtmann et al., 2016).

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; 2 simulation studies in McNeish, 2016b; van de Schoot et al., 2015; Depaoli, 2012; Holtmann et al., 2016; Miočević et al., 2017). 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, 2016b). 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; Depaoli, 2013; Holtmann et al., 2016). Depaoli (2012; 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; van de Schoot et al., 2015; Yuan & MacKinnon, 2009; 2 simulation studies in McNeish, 2016b; Holtmann et al., 2016; Miočević et al., 2017) 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. (2018) 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 3 on p. 16 in Zondervan-Zwijnenburg et al., 2018). These conclusions support the results shown in Figure 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., 2018), 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, 2016b).

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.

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, 2016a; 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, 2016a; van Erp et al., 2018). For example, Lee and 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 (2016a) 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 and Draper (2000) summarize that Bayesian estimation [BayesD and BayesN; referred to as "gently data-determined prior", and two "diffuse Inverse Wishart priors", respectively, in Appendix Table A1] 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,

2016a; van Erp et al., 2018). In the study of McNeish (2016a), BayesD (referred to as "data-dependent prior" in Appendix Table A1) is favored over the two BayesN priors (referred to as "noninformative proper/ improper Inverse Wishart priors" in Appendix Table A1), because it resulted in lower levels of bias and because of its ease of implementation. Browne and Draper (2000) report that both BayesD as BayesN (referred to as "gently data-determined prior" and two "diffuse Inverse Wishart priors" in Appendix Table A1) 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 Appendix Table A1] 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 Appendix Table A1] and Bayesian estimation using data dependent priors [BayesD; referred to as "data-driven informative priors" in Appendix Table A1] 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.

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.

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, Sijbrandij, Depaoli, Winter, Olff and van Loey (2018) and Zondervan-Zwijnenburg, Peeters, Depaoli, and van de Schoot (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 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):

$$M = i_2 + aX + e_2, \quad (1)$$

and

$$Y = i_3 + c'X + bM + e_3, \quad (2)$$

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 7, and Appendix A2 contains the R-code to reproduce the prior distributions.

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 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), 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), 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 7A).

When selecting a prior for regression coefficient *a* (see Equation 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 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.

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 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 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).

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., 2013, p. 130). To illustrate, we use this method to construct the prior distribution for the residual variance of Y, $\sigma_{e_3}^2$ (see Equation 2). Suppose a researcher collects pilot data from 20 participants and fits the mediation model, obtaining an estimate for the residual variance $\sigma_{e_3}^2$ 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 (α = 10, β = 20). This Inverse Gamma distribution can now be used as a prior distribution for residual variance $\sigma_{e_3}^2$ (see Figure 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 (α = 5, β = 10) prior distribution with smaller hyperparameters, and therefore a less informative Inverse Gamma distribution.

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), 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.

References

References marked with an asterisk (*) indicate studies included in the systematic review.

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

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

* 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. https://doi.org/10.1037/a0030642

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

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

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

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

* 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*, *21*(2), 253–262. https://doi.org/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).

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. https://doi.org/10.1037//0021-843X.111.3.425

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

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/a0031609

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

* 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.

Depaoli, S., & van de Schoot, R. (2017). Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist. *Psychological Methods*, *22*(2), 240–261.

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. https://doi.org/10.1007/s00787-015-0788-z

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

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. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

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

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

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

* 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*. https://doi.org/10.3389/fpsyg.2014.00078

* Hox, 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.

IBM Corp. (2017). IBM SPSS Statistics for Windows (Version 25). Armonk, NY: IBM Corp.

JASP team. (2018). JASP (Version 0.9). [Computer Software].

Jeffreys, H. (1945). An invariant form for the prior probability in estimation problems. Retrieved April 23, 2016, from http://rspa.royalsocietypublishing.org/

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

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics, 53*(3), 983–997. https://doi.org/10.2307/2533558

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/a0036635

Kruschke, J. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). 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. https://doi.org/10.1177/1094428112457829

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

* 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.

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. https://doi.org/10.1080/10705510701301511

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med*, *6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

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

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

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

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

* McNeish, D. (2016a). Using Data-Dependent Priors to Mitigate Small Sample Bias in Latent Growth Models A Discussion and Illustration Using Mplus. *Journal of Educational and Behavioral Statistics*, *41*(1), 27–56. https://doi.org/10.3102/1076998615621299

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

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

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. https://doi.org/10.1080/10705511.2017.1312407

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. https://doi.org/10.1371/journal.pmed.1000097

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

Natarajan, R., & Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, *95*(449), 227-237.

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

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, West Sussex, England.

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.

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. https://doi.org/10.1016/j.jclinepi.2017.04.008

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.

* 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. https://doi.org/10.1080/10705511.2014.937322

* 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. https://doi.org/10.1111/ajps.12001

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

* 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*(1). https://doi.org/10.3402/ejpt.v6.25216

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). https://doi.org/10.1080/20008198.2017.1375339

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., & van Loey, N. E. (2018). Bayesian PTSD-Trajectory Analysis with Informed Priors Based on a

Systematic Literature Search and Expert Elicitation. *Multivariate Behavioral Research*, *53*(2), 267–291. https://doi.org/10.1080/00273171.2017.1412293

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, *Psychological Methods*, *22*(2), 217–239. 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*.

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.

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yang, R., & Berger, J. O. (1996). *A catalog of noninformative priors*. Institute of Statistics and 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). https://doi.org/10.1348/000711010X497442

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

* Zondervan-Zwijnenburg, M., Depaoli, S., Peeters, M., & van de Schoot, R. (2018). Pushing the Limits: The performance of ML and Bayesian estimation with Small and Unbalanced Samples in a Latent Growth Model. *Methodology, 1*(1), 1-13.

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.

Footnotes

[1] In the current paper, 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. (2013), Kaplan (2014), Kaplan and Depaoli (2013), Kruschke (2014), Lynch (2007), Lee and Wagenmakers (2014), and for recent methodological articles to the two special issues on Bayesian Data Analysis from Psychological Methods (Chow and Hoijtink, 2017; Hoijtink and Chow, 2017).

[2] Although not further discussed in the current study, note that there are several other techniques to handle small sample sizes in SEM, such as ridge SEM (Yuan, Wu, & Bentler, 2011), and three-step estimation (Bakk, Oberski, & Vermunt, 2014).

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

[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).

[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.

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

[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.

Table 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	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 individual-level predictors (I, LS)	FIML, BayesN	M*plus* 7.1	**20, 30, 50**	4
11. McNeish, 2016a*	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 time-invariant exogenous predictors (I, LS)	FIML, REML KR, BayesN, BayesT	M*plus*, SAS PROC MIXED	**20, 30, 50**	4

				Sample Size	
Study	Model of interest	Estimation methods	Software	Number of Persons/ Clusters	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., 2018*	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., 2018*	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, 2016b*	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.

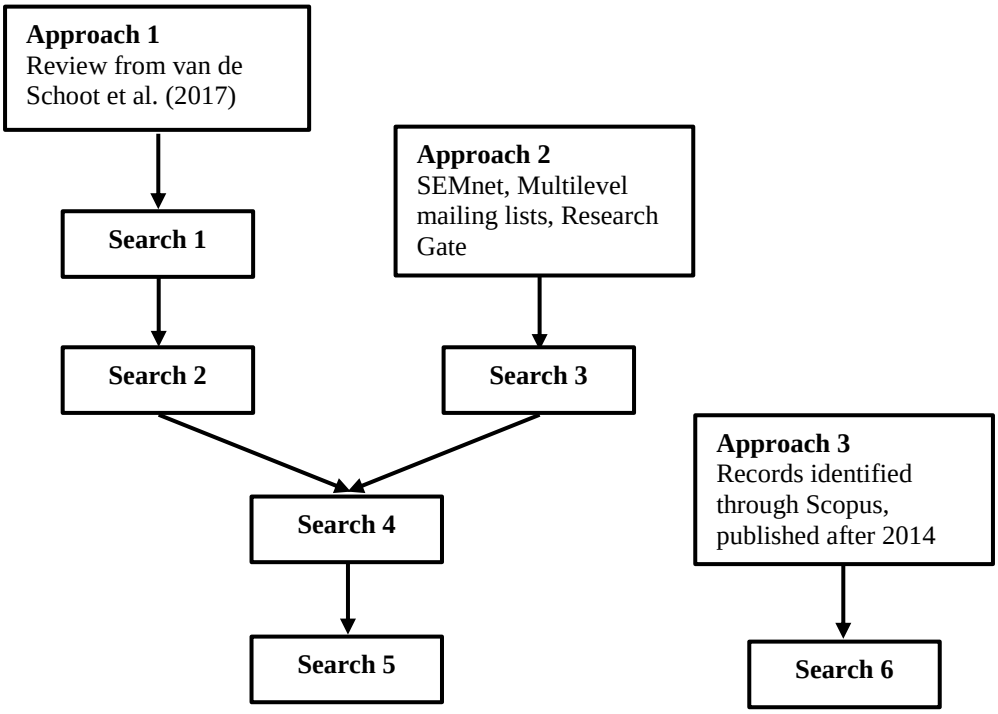

Figure 1. The three approaches and six subsequent literature searches to identify relevant references.

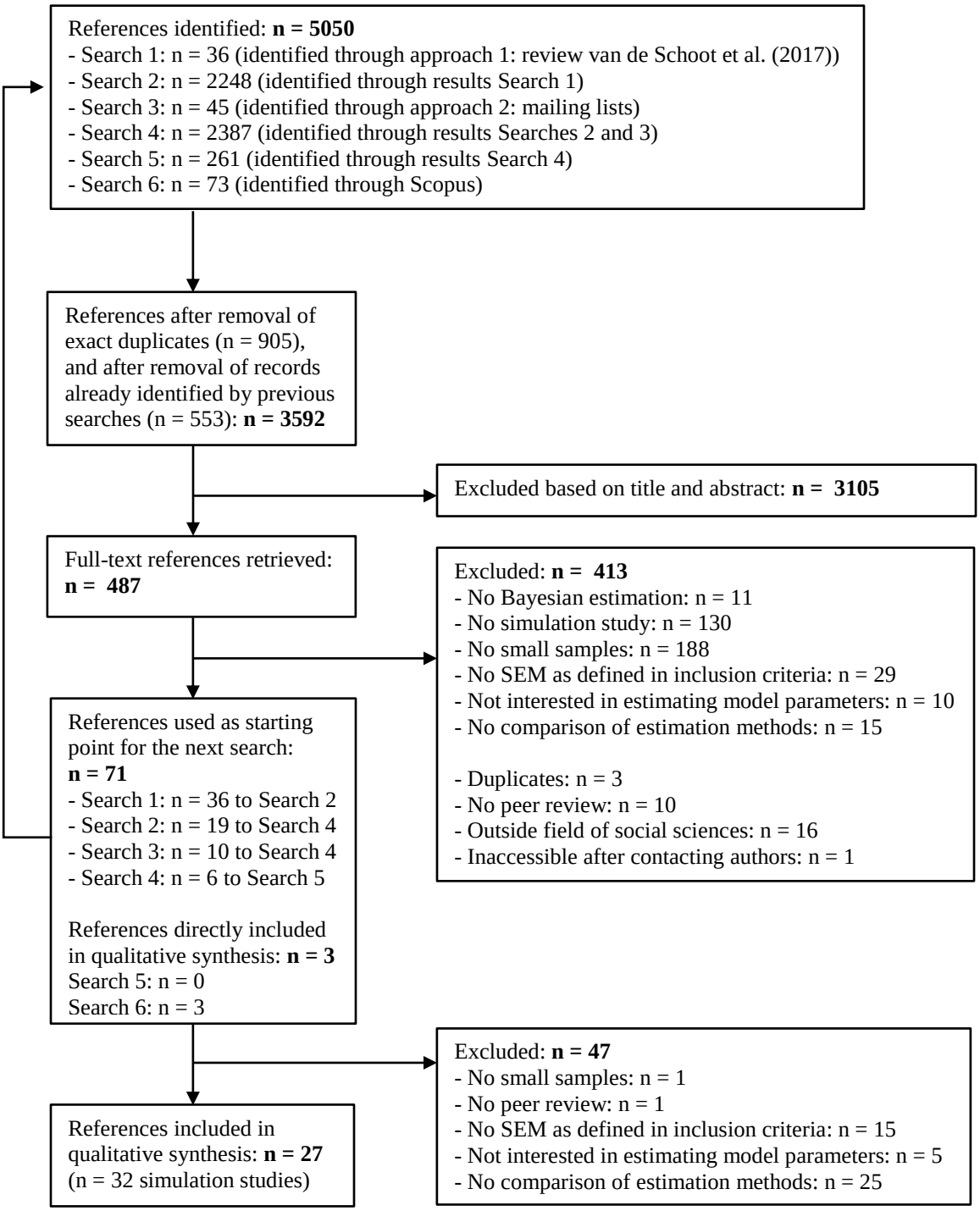

Figure 2. Summary flow chart of the search process (based on the PRISMA guidelines). For a detailed description of the exclusion criteria, we refer to the 'Inclusion and exclusion criteria' Section.

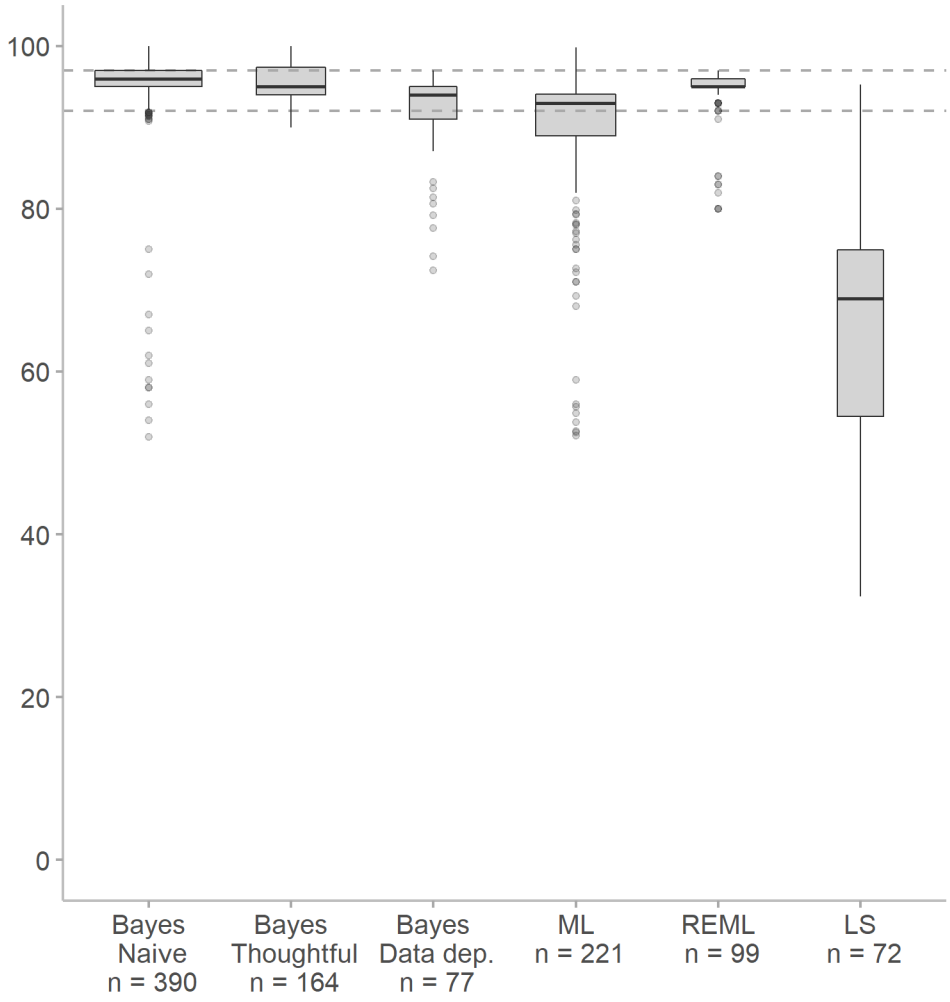

Figure 3. Reported coverage in the included studies for structural parameters (e.g., latent means, regression coefficients), for sample sizes defined as small by the original authors, presented for the varying estimation methods. Dashed grey lines represent the desirable [92.50; 97.50] coverage level 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, 2018). 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 (which is the default in ggplot2; Wickham, 2016), and the circles beyond the end of the whiskers represent outliers.

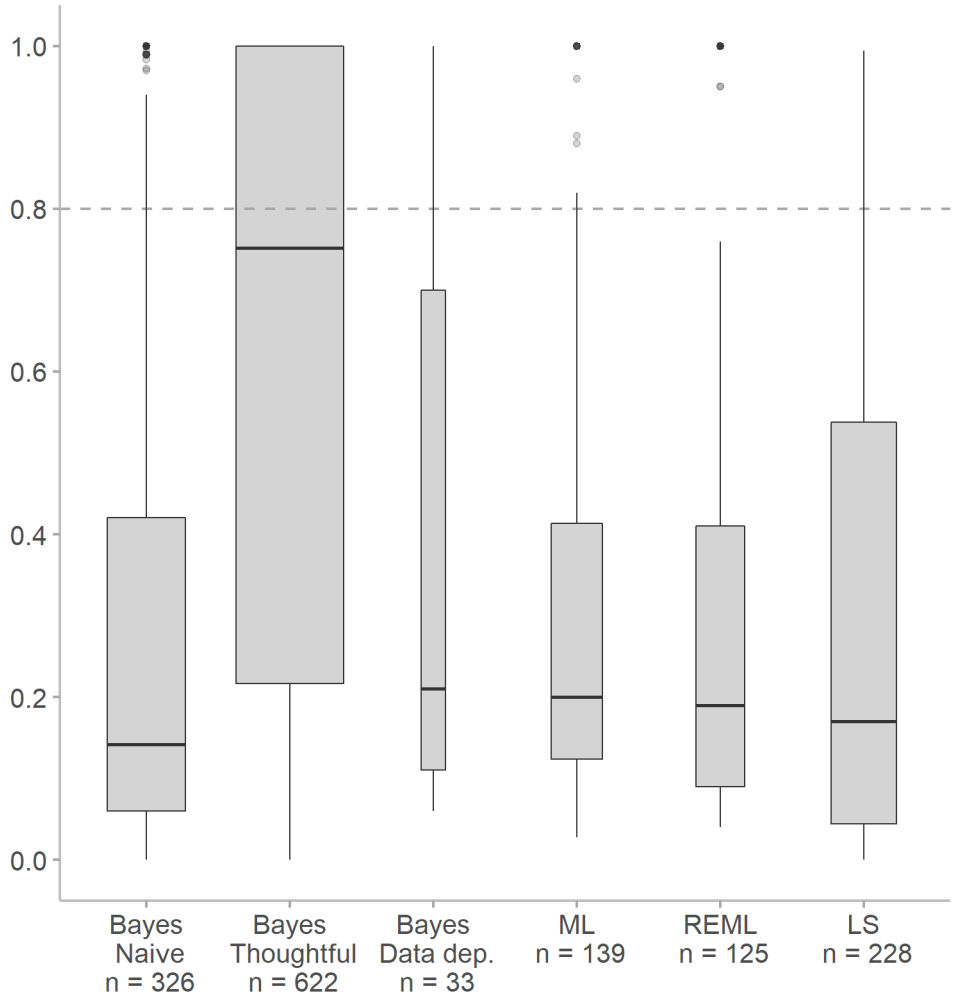

Figure 4. Reported power in the included studies for structural parameters (e.g., latent means, regression coefficients), for sample sizes defined as small by the original authors, presented for the varying estimation methods. The dashed grey line represents the desirable 0.80 power level. The *n* represents 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, 2018). 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 (which is the default in ggplot2; Wickham, 2016), and the circles beyond the end of the whiskers represent outliers.

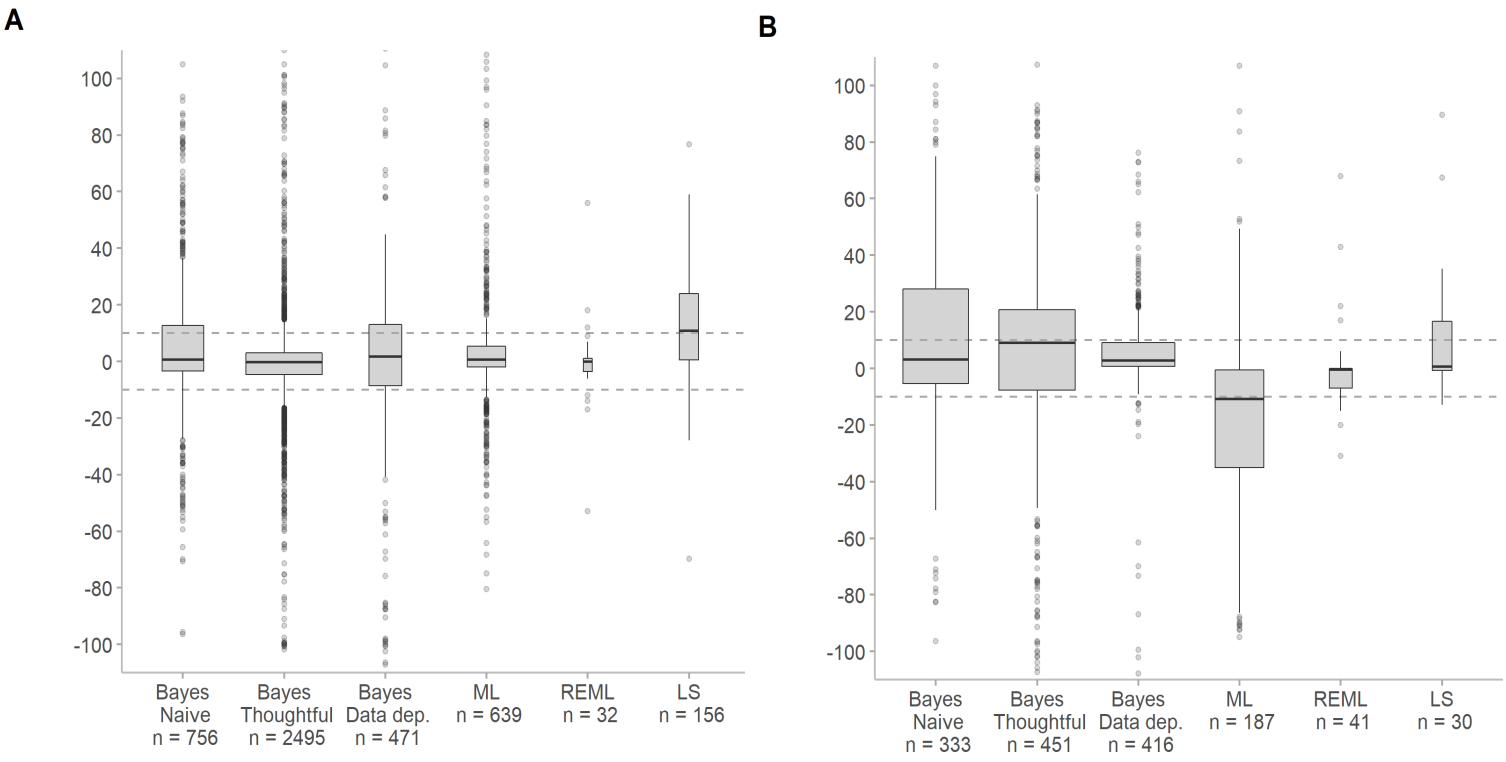

Figure 5. Reported relative bias in the included studies for A: structural parameters (e.g., latent means, regression coefficients), and B: variance parameters (e.g., factor variances, covariance, residual variances), for sample sizes defined as small by the original authors, presented for the varying estimation methods. 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, 2018). 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 (which is the default in ggplot2; Wickham, 2016), and the circles beyond the end of the whiskers represent outliers.

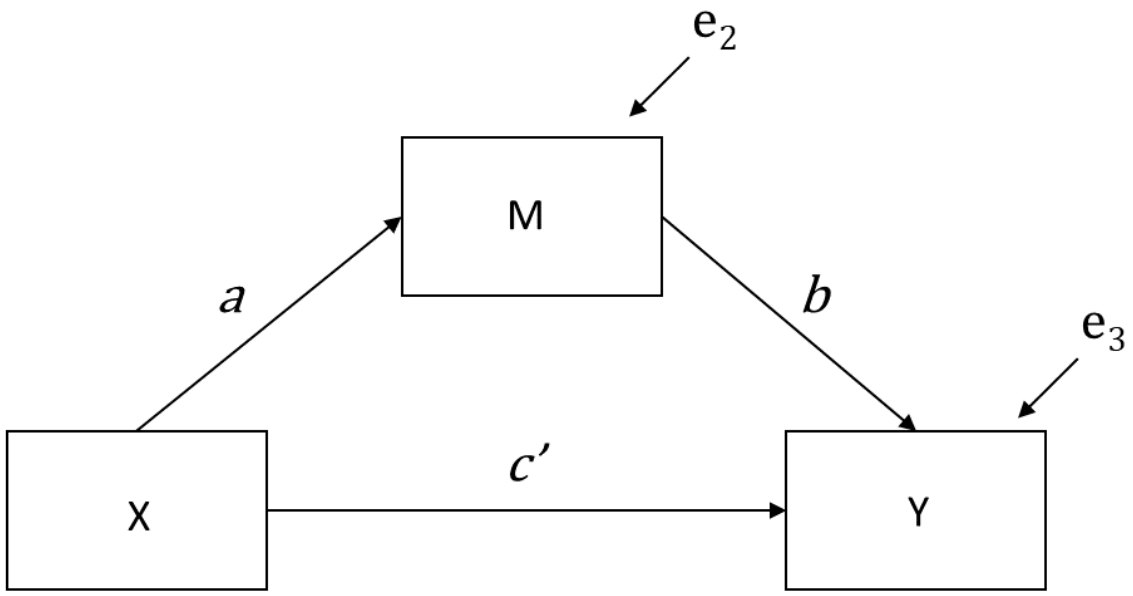

Figure 6. Single mediator model

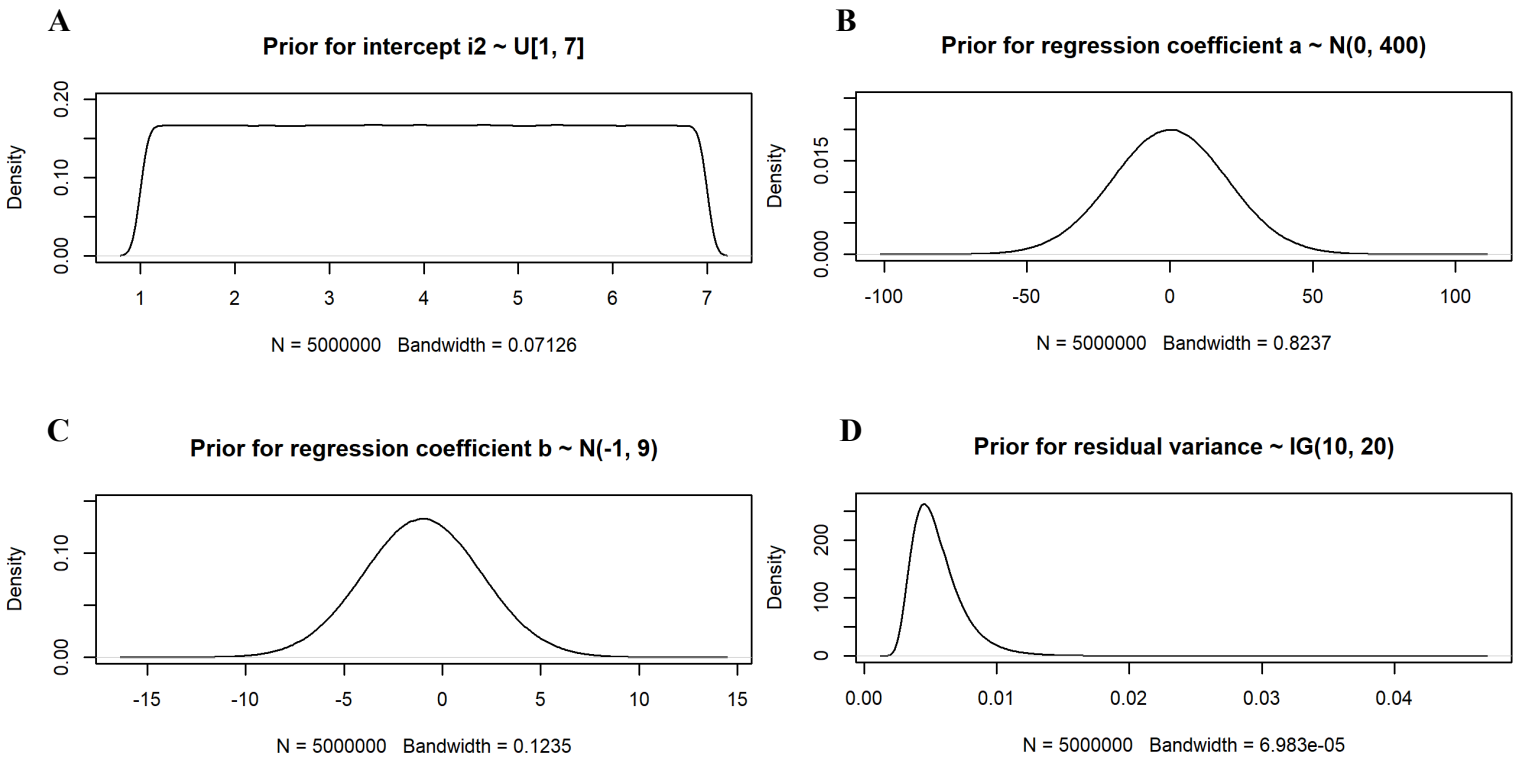

Figure 7. Uniform prior distribution for the intercept i_2 (see A), Normal prior distributions specified using mean and variance hyperparameters for regression coefficients a (see B) and b (see C) and Inverse Gamma prior distribution for the residual variance of Y specified using the shape (α) and scale hyperparameters (β; see D).

Appendix A1

Specified Prior Distributions in the Simulation Studies

Study	Wording in paper ("..."), classification in review (*)	Prior distributions
Mediation Models		
1. Chen at al., 2014	"flat or uninformative" priors *** Bayes Naive**	Intercepts of the dependent variable and mediator: τ_m $\tau_y \sim N(0, 100)$ Regression coefficients: *a,b,c,* ~ N(0, 100) Residual variables for M and Y: E_m, $E_y \sim G(0.01, 100^i)$
2. Chen at al., 2014	"flat or uninformative" priors *** Bayes Naive**	Intercepts of the dependent variable and mediator: τ_m $\tau_y \sim N(0, 100)$ Regression coefficients: $a,b,c,\lambda \sim U[-1, 1]$ Residual variables for M and Y: E_m, $E_y \sim G(0.01, 100^i)$ Residue variance of each indicator: $E_I \sim G(0.01, 100^i)$ Intercept of each indicator: $\alpha_I \sim N(0, 100)$
3. Chen et al., 2015	"informative priors that are reasonable in practice" *** Bayes Thoughtful**	First threshold: $\tau j1 \sim N(0, \tau 0)$ Other thresholds: $\Delta \tau jc$ (c = 2 to C – 1) ~ Log-N(0, $\Delta \tau 0$) Factor loadings: $\lambda j \sim N(0, \lambda 0)$ (Residual) variances: $\psi 1$, $\phi 1$, $\phi 2 \sim G(\gamma 1, \gamma 2)$ Regression paths: $\beta k \sim N(0, \beta 0)$, where $\tau 0$, $\Delta \tau 0$, $\lambda 0$, $\gamma 1$, $\gamma 2$ and $\beta 0 = 1$
	"flat priors" * **Bayes Naive**	First threshold: $\tau j1 \sim N(0, \tau 0)$ Other thresholds: $\Delta \tau jc$ (c = 2 to C – 1) ~ Log-N(0, $\Delta \tau 0$) Factor loadings: $\lambda j \sim N(0, \lambda 0)$ (Residual) variances: $\psi 1$, $\phi 1$, $\phi 2 \sim G(\gamma 1, \gamma 2)$ Regression paths: $\beta k \sim N(0, \beta 0)$, where $\tau 0$, $\Delta \tau 0$, $\lambda 0$, $\beta 0 = 100$ and $\gamma 1$, $\gamma 2 = 0.01$
4. Koopman et al., 2015	"uninformative" priors *** Bayes Naive**	Priors Model 2: Y ~ M + X. Intercept ~ N (0, 10^6) Relationships between Mediator and Outcome; Dependent and Outcome; and Covariates and Outcome ~ N (0, 10^6) Priors Model 3: M ~ X. Intercept ~ N (0, 10^6) Relationships between Dependent and Mediator; and Covariates and Mediator ~ N (0, 10^6) No information available on residual variances.
5. Miočević et al., 2017	"method of coefficients with diffuse" priors *** Bayes Thoughtful**	Regression coefficients a, b, c' ~ N(mean = population value, var = 1000) Residual variances σ_2 and $\sigma_3 \sim$ IG(.01, .01)

	"method of coefficients with informative" priors *** Bayes Thoughtful**	Regression coefficients a, b, c' ~ N(mean = population value, sd = "standard errors of respective coefficients calculated using simulated (population-generating) values at a given sample size") Residual variances σ_2 and σ_3 ~ IG(.01, .01)
	"method of covariances with diffuse" priors *** Bayes Naive**	Means of X, M, and Y ~ MVN with means of 0, variances of 1000 and covariances of 0. Covariance matrix ~ IW(df, S), where df = 3, and S was specified so "that the prior expectations for variances and covariance between variables were 1 and 0, respectively".
	"method of covariances with informative" priors *** Bayes Thoughtful**	Means of X, M, and Y ~ MVN with means of 0, variances of 1000 and covariances of 0. Covariance matrix ~ IW(S, df), where df equals the "size of the observed sample in the condition", and S was specified so "that the prior expectations for each variance and covariance equal their respective true values".
6. Yuan & MacKinnon, 2009	"informative" uniform priors *** Bayes Thoughtful**	Priors on regression parameters α, β, τ' for the varying effect sizes: α, β, τ' ~ U[-0.14, 0.14] for zero effect size α, β, τ' ~ U[0, 0.39] for small effect size α, β, τ' ~ U[0.14, 0.59] for medium effect size α, β, τ' ~ U[0.39, 0.79] for large effect size "For other parameters that appear in the mediation regression equations, noninformative prior distributions were used to reflect relatively weaker prior information on these parameters."
	"informative" normal priors *** Bayes Thoughtful**	Priors on regression parameters α, β for the varying effect sizes: N(μ, σ^2), where μ = 0, 0.14, 0.39, 0.59 for zero effect size, small effect size, medium effect size, large effect size respectively. σ^2 = not given. No prior specification given for τ'.
	"noninformative" priors *** Bayes Naive**	α, β, τ' ~ N(0, 1.0e-6[p]) Residual precision y and m: G(0.001 ,0.001)[p]
CFA Models		
7. Natesan, 2015	"informative" priors *** Bayes Thoughtful**	Correlation latent factors ξ ~ U [0, 1] Factor loadings: N (0,1) I (0,) *Truncated normal distribution, restricted to be positive.* Thresholds for category *c* and item *i:* N(0,1), where $b_{i,c-1} < b_{i,c}$ Vector of two factor scores for person *p* with mean vector mu: ω_{2p}~ N_2(mu, ∑), where mu_j~ N(0,1) and ∑ = $\begin{bmatrix}1 & \xi \\ \xi & 1\end{bmatrix}$
	"relatively less informative" priors *** Bayes Thoughtful**	Correlation between latent factors ξ ~ U [-1, 1] Factor loadings: N (0, 1) I (0,) *Truncated normal distribution, restricted to be positive* Thresholds for category *c* and item *i:* N(0,1), where $b_{i,c-1} < b_{i,c}$ Vector of two factor scores for person *p* with mean vector mu: ω_{2p}~ N_2(mu, ∑), where mu_j~ N(0,1) and ∑ = $\begin{bmatrix}1 & \xi \\ \xi & 1\end{bmatrix}$
8. Lee & Song, 2004	"data dependent priors" *** Bayes Data dependent**	"We first conducted an auxiliary Bayesian estimation on the basis of non-informative prior distribution to a single simulated data set (with n = 5a), then we chose the hyper-parameter values from the solution of this preliminary estimation. The selected hyper-parameter values are used in all the replications."
9. Van Erp et al., 2018	3 "noninformative improper" priors	Three priors for latent variable variances and residual variances: IG (0, 0), IG (-0.5, 0), IG (-1, 0) Intercepts, means, loadings, and regression coefficients: N (0, 10^{10})

	*** Bayes Naive**	
	3 "vague proper" priors *** Bayes Naive**	Three priors for latent variable variances and residual variances: IG (0.001, 0.001), IG (0.01, 0.01), IG (0.1, 0.1) Intercepts, means, loadings, and regression coefficients: N (0, 10^{10})
	"vague normal prior" *** Bayes Naive**	For measurement intercepts: N (0, 1000) For factor loadings, structural intercepts, structural regression coefficients: N (0, 100) For latent variable variances and residual variances: IG (-1, 0)
	2 "Empirical Bayes" estimators *** Bayes Data dependent**	Two priors for latent variable variances and residual variances: IG (0.5, $\hat{\sigma}^2 \cdot Q^{-1}$ $(0.5, 0.5)$), with "$\hat{\sigma}^2$ denoting the ML estimate of the variance parameter and Q^{-1} denoting the regularized inverse Gamma function." π $(\sigma^2) \propto 1$, which equals IG (-1, 0) For intercepts, means, loadings, and regression coefficients: N (0, $\hat{\mu}^2 + \hat{\sigma}^2$), where $\hat{\mu}^2$ is the squared maximum likelihood estimate, and $\hat{\sigma}^2$ is the residual variance estimate of the corresponding parameter.
Latent Growth Models		
10. McNeish, 2016a	"noninformative" improper Inverse Wishart prior *** Bayes Naive**	Factor covariance matrix: IW (0, –*p* –1) M*plus* default priors on other parameters
	"noninformative" proper Inverse Wishart prior *** Bayes Naive**	Factor covariance matrix: IW (I, *p*), where I is an identity matrix of dimension *p* M*plus* default priors on other parameters
11. McNeish, 2016a	"noninformative" improper inverse Wishart prior *** Bayes Naive**	Factor covariance matrix: IW (0, –*p* –1) M*plus* default priors on other parameters
	"noninformative" proper inverse Wishart prior *** Bayes Naive**	Factor covariance matrix: IW (I, *p*), where I is an identity matrix of dimension *p* M*plus* default priors on other parameters
	"data dependent prior" *** Bayes Data dependent**	Factor covariance matrix: FIML estimates of the (co)variances M*plus* default priors on other parameters
12. McNeish, 2016b	"improper inverse Wishart" prior *** Bayes Naive**	Growth parameter covariance matrix: IW (0, –*p* –1), "where *p* is the dimension of the covariance matrix" Residual variances: IG (-1, 0)
	"non-informative" marginal inverse Gamma prior *** Bayes Naive**	Growth parameter variances and residual variances: IG (0.01, 0.01) Growth parameter covariance: N (0, 10000)
	"weakly informative" priors * **Bayes Thoughtful**	Growth parameter covariance matrix: IW $\left(\begin{matrix}3 & 0\\ 0 & .3\end{matrix}, 6\right)$
	"strong" priors * **Bayes Thoughtful**	Growth parameter covariance matrix: IW $\left(\begin{matrix}22 & 0\\ 0 & 2.2\end{matrix}, 25\right)$

13. Van de Schoot et al., 2015	M*plus* "default" priors *** Bayes Naive**	Variance parameters: IG (-1, 0) Structural parameters: N (0, 10^{10})
	varying inverse Gamma priors *** Bayes Naive**	Variance parameters (after accounting problems with the variance parameters): IG (0, 0), IG (0.001, 0.001) Structural parameters: N (0, 10^{10})
	"very informative" Inverse Gamma prior *** Bayes Thoughtful**	Variance parameters (after accounting problems with the variance parameters): IG (0.5, 0.5) Structural parameters: N (0, 10^{10})
	"very informative" Inverse Gamma prior and prior for regression coefficient *** Bayes Thoughtful**	Variance parameters: IG (0.5, 0.5) Regression coefficient from covariate to linear slope: N (10, σ_0^2), where σ_0^2 = 10^{10}, 1000, 100, 50, 20, 10, 5, 3, 1 ("noninformative" to "very informative" prior for regression coefficient)
14. Zondervan-Zwijnenburg et al., 2018	"very informative" to "uninformative" prior for factor means *** Bayes Thoughtful**	Latent growth factor means for the reference and focal groups: N (μ_0, σ_0^2), where μ_0 = population values and σ_0^2 = 0.1, 0.3, 0.5, 1.0, 2.0, 5.0, 10^{10} M*plus* default priors for other parameters: Mean of covariate and regression coefficients: N(0, 10^{10}) Variance of covariate and residual variances of observed variables: IG(-1,0) Covariances and residual variances of growth factors: IW(0,-4)
15. Zondervan-Zwijnenburg et al., 2018	unbalanced prior information: "(…) a substantial amount of prior information (σ_0^2 = 0.1) could only be obtained for the reference group, but not for the focal group (σ_0^2 = 10.0)." *** Bayes Thoughtful**	Latent growth factor means for reference group: N(μ_0, σ_0^2), where μ_0 = population values and σ_0^2 = 0.1 ("substantial amount of prior information" Latent growth factor means for focal group: N(μ_0, σ_0^2), where μ_0 = population values and σ_0^2 = 10.0 M*plus* default priors for other parameters: Mean of covariate and regression coefficients: N(0, 10^{10}) Variance of covariate and residual variances of observed variables: IG(-1,0) Covariances and residual variances of growth factors: IW(0,-4)
Multilevel Models		
16. Baldwin & Fellingham, 2013	"thoughtful priors" * **Bayes Thoughtful**	Regression coefficient b_0 ~ N (3, 2.25) Regression coefficient b_1 ~ N (0, 1) Cluster effect u_j ~ N (0, σ_u^2) Cluster variance σ_u^2 ~ G (0.7, 0.098) Residual variance clustered condition: σ_{ec}^2~ G (13, 0.03) Residual variance unclustered condition: σ_{eu}^2 ~ G (9, 0.03)
	"flat uniform prior" instead of gamma prior for 1 cell in simulation design * **Bayes Thoughtful**	Regression coefficient b_0 ~ N (3, 2.25) Regression coefficient b_1 ~ N (0, 1) Cluster effect u_j ~ N (0, σ_u^2) Cluster variance σ_u^2 ~ U [0, 0.23]

		Residual variance clustered condition: σ^2_{ec}~ U [0, 0.69] Residual variance unclustered condition: σ^2_{eu} ~ U [0, 0.69]
17. Browne & Draper, 2000	"diffuse Inverse Wishart prior" with identity matrix * **Bayes Naive**	Fixed effects ~ U [-∞, +∞] Level 1 variance σ^2_e ~ U [0, 1/ϵ] or IG (ϵ, ϵ), where ϵ = 0.001 (The authors state these priors are equivalent.) Covariance matrix ~ IW (2, I2), where I2 is a 2 x 2 identity matrix
	"gently data-determined prior" *** Bayes Data dependent**	Fixed effects ~ U [-∞, +∞] Level 1 variance σ^2_e ~ U [0, 1/ϵ] or IG (ϵ, ϵ) where ϵ = 0.001 (The authors state these priors are equivalent.) Covariance matrix ~ IW (4, $\sum u$), where $\sum u$ is the RIGLS estimate of the covariance matrix
	"diffuse Inverse Wishart prior" * **Bayes Naive**	Fixed effects ~ U [-∞, +∞] Level 1 variance σ^2_e ~ U [0, 1/ϵ] or IG (ϵ, ϵ), where ϵ = 0.001 (The authors state these priors are equivalent.) Covariance matrix ~ IW (-3,0)
18. Browne & Draper, 2006	"diffuse" Inverse Gamma prior * **Bayes Naive**	Improper Uniform priors are used "on the real line R for fixed effects (these are functionally equivalent to proper Gaussian priors with huge variances)." Random-effect variances σ^2 ~ IG (ε, ε), where ε = 0.001
	"diffuse" Uniform prior * **Bayes Naive**	Improper Uniform priors are used "on the real line R for fixed effects (these are functionally equivalent to proper Gaussian priors with huge variances)." Random-effect variances ~ U [0, 1/ϵ], where ε = 0.001
19. Depaoli & Clifton, 2015	"noninformative (diffuse)" priors * **Bayes Naive**	Regression paths between and within level: N (0, 10^{10}) for continuous indicators and N (0, 5) for categorical indicators Variance parameters within and between level: IG (-1, 0)
	6 "weakly informative" priors * **Bayes Thoughtful**	Variance parameter within-group: IG (−1, 0) 3 levels of informativeness were specified for the regression priors: N (1, 1), N (1, 0.5), or N (1, 0.25) 2 priors for variance parameters at cluster level: IG (−1, 0) or IG (0.001, 0.001) 3 x 2 = 6 prior combinations
	"informative" prior * **Bayes Thoughtful**	Variance parameter within-group: IG (−1, 0) Regression paths between and within: N (1, 0.1) Variance parameters at cluster: IG (0.001, 0.001)
20. Farrell & Ludwig, 2008	"relatively noninformative priors" * **Bayes Thoughtful**	First-level priors on parameters: Mean of the Gaussian: μ_i ~ N (Υ_1, Υ_2), with mean Υ_1 and precision Υ_2 Standard deviation of the Gaussian in terms of precision: $1/\sigma^2 = \emptyset_i$ ~ G (δ_1, δ_2)^p Scale of exponential τ: $1/\tau = \lambda_i$ ~ G(ε_1, ε_2) Second-level priors on parameters of parent distributions: Υ_1 ~ N (0.2, 2), Υ_2 ~ G (0.5, 0.5)^p δ_1 ~ Exp (1), δ_2 ~ G (0.1, 0.1)^p ε_1 ~ Exp (1), ε_2 ~ G (0.1, 0.1)^p

21. Holtmann et al., 2016	"diffuse" priors *** Bayes Naïve**	Continuous indicator model using M*plus*: λ_{Tik}, λ_{Mik} and μ_{ik} ~ N(0, 10) Var(ϵ_{rtik}) ~ IG(-1, 0) Var(T_k), Var(M_k), Cov(T_1, T_2), Cov(M_1, M_2) ~ IW(0, -3) Categorical indicator model using M*plus*: λ_{Tik} and λ_{Mik} ~ N(0, 5) κ_{sik} ~ N(0, 10) Var(T_k), Var(M_k) ~ IW(1, 3) Cov(T_1, T_2), Cov(M_1, M_2) ~ IW(0, 3)
	"strongly informative accurate" priors * **Bayes Thoughtful**	Continuous indicator model using M*plus*: λ_{Tik} ~ N(0.8, 0.01) λ_{Mik} ~ N(1.2, 0.01) Var(ϵ_{rtik}) ~ IG(-1, 0) μ_{ik} ~ N(3, 1) Var(T_k), Var(M_k), Cov(T_1, T_2), Cov(M_1, M_2) ~ IW(0, -3) Categorical indicator model using M*plus*: λ_{Tik} ~ N(0.8, 0.01) λ_{Mik} ~ N(1.2, 0.01) κ_{sik} ~ N(0, 10) Var(T_k), Var(M_k) ~ IW(1, 3) Cov(T_1, T_2), Cov(M_1, M_2) ~ IW(0, 3)
	"weakly informative accurate" priors * **Bayes Thoughtful**	Extended set of prior conditions for categorical indicator model using M*plus*: λ_{Tik} ~ N(0.8, 0.2) λ_{Mik} ~ N(1.2, 0.2) Var(T_k), Var(M_k) ~ IW(1, 3) Cov(T_1, T_2), Cov(M_1, M_2) ~ IW(0, 3) κ_{sik} ~ N(0, 10)
	"weakly informative inaccurate" priors * **Bayes Thoughtful**	Extended set of prior conditions for categorical indicator model using M*plus*: λ_{Tik} ~ N(1.2, 0.2) λ_{Mik} ~ N(0.8, 0.2) Var(T_k), Var(M_k) ~ IW(1, 3) Cov(T_1, T_2), Cov(M_1, M_2) ~ IW(0, 3) κ_{sik} ~ N(0, 10)
	"strongly informative inaccurate" priors * **Bayes Thoughtful**	Extended set of prior conditions for categorical indicator model using M*plus*: λ_{Tik} ~ N(1.2, 0.01) λ_{Mik} ~ N(0.8, 0.01)

		$Var(T_k)$, $Var(M_k)$ ~ IW(1, 3) $Cov(T_1, T_2)$, $Cov(M_1, M_2)$ ~ IW(0, 3) κ_{sik} ~ N(0, 10)
	"informative Wishart" prior * **Bayes Thoughtful**	Extended set of prior conditions for categorical indicator model using M*plus*: λ_{Tik} and λ_{Mik} ~ N(0, 5) $Var(T_k)$, $Var(M_k)$ ~ IW(123, 126) $Cov(T_1, T_2)$ ~ IW(61.5, 126) $Cov(M_1, M_2)$ ~ IW(49.2, 126) κ_{sik} ~ N(0, 10)
22. Hox, van de Schoot, Matthijsse, 2012	"uninformative priors" * **Bayes Naive**	M*plus* (version 6.1) default priors are used. No prior distributions are described in the paper.
23. Hox et al., 2014	"flat" M*plus* default priors * **Bayes Naive**	Path coefficients ~ N (0, 10^{10}) Variances ~ IG (-1,0)
24. McNeish, 2016b	"non-informative M*plus* default" priors * **Bayes Naive**	Random intercept variance: IG (-1, 0) For all other parameters: M*plus* default priors are used.
	"non-informative" prior * **Bayes Naive**	Random intercept variance: IG (0.01, 0.01) For all other parameters: M*plus* default priors are used.
	"weakly informative" priors * **Bayes Thoughtful**	Random intercept variance: IG (3, 3.25), which results in a mean of 1.625 and a standard deviation of 1.625. For all other parameters: M*plus* default priors are used.
	"strongly informative" priors * **Bayes Thoughtful**	Random intercept variance: IG (12, 18), which results in a mean of 1.64 and a standard deviation of 0.52. For all other parameters: M*plus* default priors are used.
25. McNeish & Stapleton, 2016	"uninformative" IG prior * **Bayes Naive**	Variance components ~ IG (0.01, 0.01)
	"uninformative" U prior * **Bayes Naive**	Variance components ~ U [0, 10]
	"uninformative" Half-Cauchy prior * **Bayes Naive**	Variance components ~ Half-Cauchy (0, 4)
26. Stegmueller, 2013	"non-informative, vague prior" * **Bayes Naive**	Level 1 and level 2 variances: IG (0.001, 0.001) In models containing a random coefficient: Variance covariance matrix: IW (S, *d*), with d degrees of freedom, and diagonal scale matrix S = I_2. (This prior produces "a marginal prior for the correlation between intercept and slope, which is uniform on [-1, 1], and distributed IG (1, ½) for the two variances.")
	"non-informative, vague prior" * **Bayes Naive**	Level 1 variances: IG (0.001, 0.001) Level 2 variances: "distributed uniform on the standard deviation" $\sqrt{\sigma_y^2}$ ~ c In models containing a random coefficient: Variance covariance matrix: Either IW (S, *d*), with d degrees of freedom, and diagonal scale matrix S = I_2, or an "IW prior which posits twice the size for the (diagonal) variances" is used.

27. Tsai & Hsiao, 2008	"reference" approximate Uniform Shrinkage prior * **Bayes Naive**	Fixed effects: N (0, 10^6) Covariance matrix D: reference approximate Uniform Shrinkage prior $\pi_{us}(D) \propto \det\left(\mathrm{I}_q + \left\{\frac{1}{l}\sum_{i=1}^{I} Z_i^t W_i Z_i\right\} D\right)^{-q-1}$, Where I_q is a *q x q* identity matrix, and W_i is an n_i *x* n_i diagonal weight matrix with elements $1 \Big/ \left\{ v_{ij}^b \left[\partial \eta_{ij}^b \Big/ \partial \mu_{ij}^b \right]^2 \right\}$
	"reference" Jeffreys prior using Fisher information matrix * **Bayes Naive**	Fixed effects: N (0, 10^6) Covariance matrix D: Jeffreys' prior using the approximate Fisher information matrix $\mathrm{I}(\theta)$ $\left(= \mathrm{I}\left(\mathrm{D}(\theta)\right)\right)$ for D with the (*j, k)*th component $\mathrm{I}_{jk} = \frac{1}{2} tr\left(\mathrm{P} \frac{\partial \mathrm{V}}{\partial \theta_j} \mathrm{P} \frac{\partial \mathrm{V}}{\partial \theta_k}\right)$, Where $\mathrm{P} = \mathrm{V}^{-1} - \mathrm{V}^{-1}\mathrm{X}(\mathrm{X}^t\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{X}^t\mathrm{V}^{-1}$ and $\mathrm{V} = \mathrm{W}^{-1} + \mathrm{ZDZ}^t$
	"reference" Jeffreys prior derived from different approximate likelihood * **Bayes Naive**	Fixed effects: N (0, 10^6) Covariance matrix D: Jeffreys' prior, as suggested by Natarajan & Kass (2000), derived from approximate likelihoods $\mathrm{I}_{jk}^{*} \approx \sum_{i=1}^{I} tr\left(([\mathrm{Z}_i^t \mathrm{W}_i \mathrm{Z}_i]^{-1} + \mathrm{D})^{-1} \frac{\partial \mathrm{D}}{\partial \theta_j} \times ([\mathrm{Z}_i^t \mathrm{W}_i \mathrm{Z}_i]^{-1} + \mathrm{D})^{-1} \frac{\partial \mathrm{D}}{\partial \theta_k}\right)$
AR Model		
28. Price, 2012	"informative priors" *** Bayes Thoughtful**	For theta ~ MVN (0,4) Covariance matrix ~ IW(η_0, S_0^{-1})
Mixture Models		
29. Depaoli, 2012	"weak" priors *** Bayes Thoughtful**	Factor loadings: N(μ, 100), where μ = population value Default M*plus* prior for class proportions: D(10, 10) For all other parameters M*plus* default priors are used.
	"tight" priors * **Bayes Thoughtful**	Factor loadings: N(μ, 0.01), where μ = population value Class proportions ~ Dirichlet prior reflecting sample size and mixture class proportions, for n = 100, class proportion 0.80/0.20, and prior is D(80,20) For all other parameters M*plus* default priors are used.
30. Depaoli, 2012	"weak" priors *** Bayes Thoughtful**	Factor loadings: N(μ, 100), where μ = population value Default M*plus* prior for class proportions: D(10, 10) For all other parameters M*plus* default priors are used.
	"tight" priors * **Bayes Thoughtful**	Factor loadings: N(μ, 0.01), where μ = population value Class proportions ~ Dirichlet prior reflecting sample size and mixture class proportions, for n = 100, class proportion 0.80/0.20, and prior is D(80,20) Factor means ~ N(μ, 0.10), where μ = population value Factor variances and covariances ~ IW(Ω, *d*), where Ω = population value, *d* = dimension of variance-covariance matrix plus 1
31. Depaoli, 2013	M*plus* "default noninformative" priors *** Bayes Naive**	M*plus* default priors on all parameters Default prior for class proportions: D(10,10,10)

	"informative accurate" priors * **Bayes Thoughtful**	Growth parameters: N(μ, σ^2), where μ = population value, and σ^2 = 5% of corresponding population value. Class proportions: Dirichlet prior with values indicating accurate knowledge of class sizes: for n = 150 and class proportions: 0.33/0.33/0.33, the prior is D(50,50,50), 0.45/0.45/0.10, the prior is D(67,67,15), 0.70/0.20/0.10, the prior is D(105,30,15)
	"data-driven informative" priors * **Bayes Data dependent**	Growth parameters: N(μ, σ^2), where μ and σ^2 = average of maximum likelihood parameter estimate, and average variance respectively, across 1000 replications. Class proportions: D(1,1,1)
	"weakly informative" priors * **Bayes Thoughtful**	Growth parameters: N(μ, σ^2), where μ = population value, and σ^2 = 50% of corresponding population value. Class proportions: D(1,1,1)
	"partial informative" priors * **Bayes Thoughtful**	Intercept growth parameters: N(μ, σ^2), where μ = population value, and σ^2 = 5% of corresponding population value. For class proportions and slope growth parameters, M*plus* default priors are used.
	Bayes "informative and inaccurate priors" * **Bayes Thoughtful**	Growth parameters: N(μ, σ^2), where σ^2 = fixed to 50% of corresponding population value, and μ = population value minus 3 standard deviations based on the fixed variance parameter. For class proportions, M*plus* default priors are used.
32. Serang et al., 2015	"informative" priors for the mixture proportions * **Bayes Thoughtful**	Means of b_{ji1}, b_{ji2} and b_{ji3} for class 1: N(0, 0.001)^{p} "Both the mean intercept and the mean change to the upper asymptote parameters were restricted to be greater in the second class (β21 > β11; β22 > β12)." Means of b_{ji1} and b_{ji2} for class 2: N(0, 0.001)^{p} + difference, where difference ~ U[0,100] Mean of b_{ji3} for class 2: N(0, 0.001)^{p} Variances of b_{ji1}, b_{ji2} and b_{ji3} ~ W(1,3) Covariances of b_{ji1}, b_{ji2} and b_{ji3} ~ W(0,3) Class proportions: D(160,40)

Note. Prior distributions when and as given in text or appendix of the included studies. Prior distributions are specified on the variance, unless stated otherwise. ^{i} inverse scale parameter, ^{p} = precision (inverse variance) is used. "Between quotation marks" = wording used by original authors to describe these prior distributions. **Bold =** the category the priors were assigned to in the current study (Bayes Thoughtful, Bayes Naive, Bayes Data dependent). Abbreviations prior distributions: D = Dirichlet, Exp. = Exponential, G = Gamma, IG = Inverse Gamma, IW = Inverse Wishart, Log-N = Log Normal, MVN = multivariate normal, N = Normal, U = Uniform, W = Wishart. Note that although the content of the table is formatted in a standardized way, we used the original wording for the parameters and prior distributions of the studies, to keep the original authors choices intact.

Appendix A2: R-code to reproduce the prior distributions as discussed in "Recommendations on how to Construct Thoughtful Priors".

```
### Figure 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 7B, regression coefficient a
# Normal prior distribution based on expert knowledge on implausible and
      plausible values. Most implausible positive/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)
plot(density(x), main = paste("Prior for regression coefficient a ~ N(",
      mean, ", ", var,")", sep=""), ylim = c(0, 0.025))

### Figure 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 7D, residual variance parameter
# Inverse Gamma prior distribution for the residual variance parameter

library(MCMCpack)
# 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))
```